Annals of Computer Science and Information Systems
Volume 6

Position Papers of the 2015 Federated Conference on Computer Science and Information Systems

September 13–16, 2015. Łódź, Poland

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

PTI

Annals of Computer Science and Information Systems, Volume 6

Position Papers of the 2015 Federated Conference on Computer Science and Information Systems

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

2015, Warszawa,
Polskie Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 6

Position Papers of the 2015 Federated Conference on Computer Science and Information Systems

USB: ISBN 978-83-60810-77-4
WEB: ISBN 978-83-60810-76-7

ISSN 2300-5963
DOI 10.15439/978-83-60810-76-7

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

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

Cover:
Jana Waleria Denisiuk,
Elbląg, Poland

Also in this series:

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

Volume 5: Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-66-8, ISBN USB: 978-83-60810-67-5**

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

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

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

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

DEAR Reader, it is our pleasure to present to you Position Papers of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Łódź, Poland, on September 13-16, 2015. This is the third year when position papers have been introduced as a separate category of contributions. They represent emerging research papers and challenge papers. The former present preliminary research results from work-in-progress based on sound scientific approach but presenting work not completely validated as yet. The latter propose and describe research challenges in theory or practice of computer science and information systems.

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

The following FedCSIS 2015 events included position papers in their program:

- AAIA'15—10th International Symposium Advances in Artificial Intelligence and Applications
  - AIMA'15—5th International Workshop on Artificial Intelligence in Medical Applications
  - WCO'15—8th Workshop on Computational Optimization
- CSS—Computer Science & Systems
  - BCPC'15—1st International Workshop on Biological, Chemical and Physical Computations
  - CANA'15—8th Computer Aspects of Numerical Algorithms
  - IWCPS'15—2nd International Workshop on Cyber-Physical Systems
  - WAPL'15—5th Workshop on Advances in Programming Languages
- iNetSApp'15—3rd International Conference on Innovative Network Systems and Applications
  - EAIS'15—2nd Workshop on Emerging Aspects in Information Security
  - SoFAST-WS'15—4th International Symposium on Frontiers in Network Applications, Network Systems and Web Services
  - WSN'15—4th International Conference on Wireless Sensor Networks
- IT4MBS—Information Technology for Management, Business & Society
  - ABICT'15—6th International Workshop on Advances in Business ICT
  - AITM'15—13th Conference on Advanced Information Technologies for Management
  - ISM'15—10th Conference on Information Systems Management
- JAWS—Joint Agent-oriented Workshops in Synergy
  - ABC:MI'15—10th Workshop on Agent Based Computing: from Model to Implementation
  - MAS&S'15—9th International Workshop on Multi-Agent Systems and Simulations
  - SEN-MAS'15—4th International Workshop on Smart Energy Networks & Multi-Agent Systems

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

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

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

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

Annals of Computer Science and Information Systems, Volume 6

Position Papers of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS)

September 13–16, 2015. Łódź, Poland

TABLE OF CONTENTS

10TH International Symposium Advances in Artificial Intelligence and Applications

5TH International Workshop on Artificial Intelligence in Medical Applications

4TH Joint Agent-oriented Workshops in Synergy

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

9TH International Workshop on Multi-Agent Systems and Simulation

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

10th International Symposium
Advances in Artificial Intelligence and Applications

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

Topics

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

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

Professor Zdzislaw Pawlak Best Paper Awards

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

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

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

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

IFSA Award for Young Scientist

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

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

Event Chairs

Janusz, Andrzej, University of Warsaw, Poland

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

Advisory Board

Kacprzyk, Janusz, Systems Research Institute, Warsaw, Poland

Kwaśnicka, Halina, Wroclaw University of Technology, Poland

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

Skowron, Andrzej, University of Warsaw, Poland

Program Committee

Artiemjew, Piotr, University of Warmia and Mazury, Poland

Bartkowiak, Anna, Wroclaw University, Poland

Bazan, Jan, University of Rzeszów, Poland

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

Błaszczyński, Jerzy, Poznan University of Technology, Poland

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

Chakraverty, Shampa, Netaji Subhas Institute of Technology, India

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

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

Czarnowski, Ireneusz, Gdynia Maritime University, Poland

Dardzinska, Agnieszka, Bialystok University of Technology, Poland

Dey, Lipika, Tata Consulting Services, India

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

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

Girardi, Rosario, Federal University of Maranhão, Brazil

Hassanien, Aboul Ella, Cairo University, Egypt

Herrera, Francisco, University of Granada, Spain

Holzinger, Andreas, Graz University of Technology, Austria

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 3–9

DOI: 10.15439/2015F399
ACSIS, Vol. 6

Predicting Thyrotoxicosis in Patients Using a Set of Routine Tests: Adding their Rate of Annual Time-Series Variations to Self-Organizing Map-Based Predictive Model Improves Diagnostic Accuracy

Sorama Aoki, Sono Nishizaka, Kenichi Sato, Kenji Hoshi, Junko Kawakami
Medical and Pharmaceutical Information Science, Tohoku Pharmaceutical University,
4-4-1 Komatsushima, Aoba-ku, Sendai 981-8558, Japan.
Email: s-aoki@tohoku-pharm.ac.jp, 21452111@is.tohoku-pharm.ac.jp,
{ksato, hoshi, jnaka}@tohoku-pharm.ac.jp

Kouki Mori
Center for Health Promotion,
JR Sendai Hospital
Sendai, Japan.
kouki-mori@jreast.co.jp

Yoshinori Nakagawa
Sendai Thyroid Clinic
Sendai, Japan.
na@sendaikojosen.com

Wataru Hida, Katsumi Yoshida
Department of Health Supervision,
Tohoku Kosai Hospital
Sendai, Japan.
wa-hida@tohokukosai.com,
kayosimd@beetle.ocn.ne.jp

***Abstract*— Difficulties have been associated with accurately diagnosing patients with thyroid dysfunction (PTD); however, measuring thyroid hormone levels in all individuals is challenging. We successfully constructed a prediction model for PTD by adopting pattern recognition methods using a combination of six routine laboratory tests, and identified 21 new PTD using our screening method, which was executed at two health check-up centers. In the present study, we newly introduced time-series variations in routine tests as additional parameters in order to develop the model by eliminating the influence of individual differences in routine tests. We constructed self-organizing maps (SOM) using the time-series traceable data of 13 PTD and 45 healthy individuals. We then investigated the locations of 140 projected false positives in our previous study on SOM and found that the number of false positives markedly decreased, thereby demonstrating the progression of our new model.**

I. Introduction

PATIENTS with thyroid dysfunction (PTD) are often overlooked and misdiagnosed. For example, thyrotoxicoses have been misdiagnosed as a heart disease or malignant tumor of the digestive tract, while hypothyroidism has been misdiagnosed as muscle, heart, or liver disorders or hyperlipidemia in general screening by doctors or internists. Therefore, these patients frequently receive the wrong treatment [1]-[4]. Thyroid dysfunction progresses slowly as a result of PTD being overlooked or misdiagnosed in addition to patients often not being aware of and tolerating their symptoms, which results in many patients remaining unexamined.

Thyroid specialists previously reported difficulties in identifying PTD based on physical findings alone; therefore, the measurement of TSH levels is considered indispensable [5], [6]. However, TSH levels cannot be measured in all individuals who visit a clinic due to the associated costs, such that receiving a full health check-up in this regard is difficult in terms of cost-effectiveness and remains a topic of debate [7].

Since an excess or lack of thyroid hormones influences the whole body and produces abnormal routine test results, we attempted to construct a predictive model using an appropriate set of routine tests, which are measured in the health check-up system, in order to detect PTD using a low-cost method. In our previous studies, we analyzed the routine test data of PTD and healthy individuals using three types of pattern recognition methods (PRM) [8]-[10] in addition to medical statistics. We found that PRMs with three parameters (an elevation in alkaline phosphatase (ALP) and decreases in serum creatinine (S-Cr) and total cholesterol (TC)) [11]-[13], or four parameters (the previous three parameters in addition to an elevation in heart rate (HR)) [14] allowed accurate screening for hyperthyroidism (thyrotoxicosis). We also reported that PRMs with another set of four parameters (elevations in lactate dehydrogenase (LDH), S-Cr, and TC, and a lower red blood cell count (RBC)) allowed accurate screening for hypothyroidism [15]. We applied these predictive models to the screening of 4,355 Japanese people whose routine test data had already been measured in the general health check-up system, referred to as “the Ningen Dock”, at JR Sendai Hospital between July 2008 and December 2011, and identified 7 overt PTD (2

This work was supported in part by a Grant-in-Aid for Scientific Research (23590811, 26460777) from the Japan Society for the Promotion of Science.

patients with Graves' disease, 2 with painless thyroiditis, and 3 with hypothyroidism), who were subsequently treated by thyroid specialists [14], [16]. None of the 7 PTD had expressed concerns regarding their health. We also applied these predictive models to the screening of 8,831 examinees of the Ningen Dock at Tohoku Kosai Hospital, which is a larger scale institution than JR Sendai Hospital, between September 2011 and March 2013, and successfully identified 14 overt PTD (8 patients with Graves' disease, 2 with painless thyroiditis, and 4 with hypothyroidism) [17]. Although we identified 21 undetected PTD, there were 218 false positives (91 at JR Sendai Hospital and 127 at Tohoku Kosai Hospital) in our screening who were subsequently found to be normal after measuring serum levels of thyroid hormones (free thyroxin (FT4) and thyroid stimulating hormone (TSH)). Most of these false positives originated from the low threshold of the predicted probability (60%) in the screening to prevent false negatives. An analysis of the screening results at the Ningen Dock of Tohoku Kosai Hospital revealed that setting a higher threshold value for probabilities (85%) that satisfied the condition of not overlooking patients with Graves' diseases and with overt hypothyroidisms, who were found in the screening, decreased the number of false positives to 45 [17]. Although this threshold level caused two false negatives, both were patients with painless thyroiditis who did not require medical treatment; therefore, overlooking these cases did not yield any clinical problems.

We found that many of the time-series variations in the routine tests of false positives did not always change toward a pattern characteristic of a thyroid dysfunction. In order to quantitatively evaluate these time-series changes, we parsed time-series variations by calculating the average rates of the annual time-series variation (RATV) in each of the routine test data between previous and current visitations to the Ningen Dock. We only selected cases suspected of thyrotoxicosis in our screening with a previous visitation record to the Ningen Dock of within three years. In each time-series traceable subject, we significantly decreased the number of false positives from 140 to nine, while maintaining the true positives of Graves' disease, by simultaneously plotting both the predicted probability and average RATV in a scattergram [18].

In the present study, we attempted to construct an advanced predictive model with PRMs, especially Self-Organizing maps (SOM), by introducing the 4 RATVs of routine tests (ALP, S-Cr, TC, and HR) in subjects suspected of having thyrotoxicosis in our screening to analyze RATV more non-linearly than our former analysis. We then determined whether introducing RATV into the SOM model improved diagnostic accuracy.

This study was approved by the Ethical Review Boards of Tohoku Kosai Hospital and JR Sendai Hospital. Data of all subjects were handled using linkable anonymization.

II. SUBJECTS AND METHODS

A. Subjects

At the Ningen Docks of JR Sendai Hospital (between July 2008 and December 2014) and Tohoku Kosai Hospital (between September 2011 and March 2013), we identified 13 true positives (6 at JR Sendai Hospital and 7 at Tohoku Kosai Hospital) and 140 false positives (104 at JR Sendai Hospital and 36 at Tohoku Kosai Hospital) of thyrotoxicosis in our screening, whose previous visitation records to the Ningen Dock were within three years. Furthermore, 45 healthy female controls were collected from the Ningen Dock of JR Sendai Hospital to train normal data for the SOM model; therefore, 198 subjects were parsed in the present study. In those screened, we used the predictive screening tool shown in Figure 1 and obtained the predictive probability for examinees after inputting six routine tests and the presence/absence of medication with dyslipidemic therapeutic drugs. The threshold level of the predicted probability for the need for hormone measurements was set to 60%. The mean ± SD interval days of visitation was 383±68 days for true positives, 379±79 days for false positives, and 399±100 days for healthy controls; every interval day was nearly equal to a year.

Fig. 1. The interface of our simple screening tool based on a predictive model that instantaneously yields classification results for examinees if their routine test (ALP, S-Cr, TC, LDH, RBC, and HR) data are input [17].

Table 1 shows the 13 patients (true positives) with thyrotoxicosis identified by our screening. The clinical and laboratory features of each patient shown in the table are the current visitation values. The diagnostic results of patients who consulted the Thyroid Outpatients Service of JR Sendai Hospital or Tohoku Kosai Hospital according to our encouragement after our screening are shown for hyperthyroidism as Graves' disease or destructive thyroiditis (painless or sub-acute).

B. Rates of annual time-series variations in routine tests

In order to handle each time-series variation in the routine tests quantitatively, we calculated the rates of annual time-series variations in each subject's routine test data between previous and current visitations to the Ningen Dock [18]. In this calculation, we divided the interval days of visitations in each subject by 365 days in order to normalize inconsistencies in the interval day; therefore, the rate of the annual time-series variation (RATV) was obtained by the following equation:

$$v_R = \frac{R_c - R_p}{R_p} \times \frac{1}{\frac{D_c - D_p}{365}} \quad (1)$$

where v and R denote the RATV and measured value of each routine test (ALP, TC, S-Cr, and HR), and D denotes the visitation date to the Ningen Dock. Subscript c denotes the current visitation, while p denotes the previous visitation to the Ningen Dock. Therefore, v_{ALP} , v_{TC} , v_{S-Cr} and v_{HR} were obtained for each subject.

C. Analyses using pattern recognition methods

As we reported previously [11]-[13], [15], routine test data were analyzed using three types of PRMs together in order to make our predictive model more robust. The first method, SOM [8], is a simplified model that reflects biological neural networks that realize character extraction in the cortex, and is also a non-linear extension of a principal component analysis in statistics. This method is superior in terms of classification capability and visualization of data distribution. If the SOM prepared from the samples from individuals with definite diagnoses shows clustering reflecting these diagnoses, we have the ability to predict which patients have thyrotoxicosis by determining the locations of individuals with unknown diagnoses projected on the SOM according to their routine test data. Furthermore, by analyzing the characteristics of the data within the component plane for each parameter, it is possible to select parameters that are useful for making a diagnosis [11]-[13], [15]. In the present study, we used the SOM_PAK package [19] after adding a facility of our own. We then adopted Bayesian regularized neural networks (BRNN) [9]. This is a multi-layer neural network, but was extended to include the Bayesian probability framework for treating model parameters in order to avoid defects such as overfitting, which are encountered in the conventional maximum likelihood approach. We used the Software for Flexible Bayesian Modeling package [20] by Neal for the BRNN. We then adopted the support vector machine (SVM) [10] with the LIBSVM package [21] and selected radial basis function kernels. We showed in Figure 1 the interface of our predictive screening tool [17], which displays the three types of predictive outputs for both hyperthyroidism and hypothyroidism in an examinee. The calculated probabilities by BRNN and SVM are given by %. We judged through visual observations whether each examinee had a thyroid dysfunction as well as the degree of severity by noting the location projected on the SOM (labeled by a yellow star).

In the present study, we added four RATV parameters (v_{ALP} , v_{TC} , v_{S-Cr} , and v_{HR}) to the set of four routine tests (ALP, TC, S-Cr, and HR) used to date; therefore, we adopted a set of eight parameters (ALP, TC, S-Cr, HR, v_{ALP} , v_{TC} , v_{S-Cr} , and v_{HR}) to construct the predictive model by using the SOM. We initially constructed the SOM using the training data set (13 true positives and 45 healthy controls) to determine whether these eight parameters had the ability to classify these two groups. One hundred and forty

TABLE 1.
CLINICAL AND LABORATORY FEATURES OF 13 TIME-SERIES TRACEABLE PATIENTS WITH OVERT THYROTOXICOSIS IDENTIFIED IN OUR SCREENING, WHOSE PREVIOUS VISITATION RECORDS TO THE NINGEN DOCK WAS WITHIN THREE YEARS

	Subject No.	Sex	Age	Diagnosis	FT4 (ng/dL)	TSH (μIU/mL)	ALP (IU/L)	S-Cr (mg/dL)	TC (mg/dL)	HR (/min)
Tohoku Kosai Hospital	K1	Female	41	Painless Thyroiditis	1.76*1	<0.01	216	0.47	164	81
	K2	Female	54	Graves' Disease	3.24*1	<0.01	154	0.37	131	89
	K3	Male	54	Graves' Disease	3.56*1	<0.01	376	0.65	148	70
	K4	Female	57	Graves' Disease	3.35*1	<0.01	396	0.57	134	81
	K5	Female	61	Graves' Disease	3.65*1	<0.01	375	0.33	128	58
	K6	Female	47	Graves' Disease	2.25*1	<0.01	392	0.45	146	92
	K7	Female	41	Graves' Disease	3.78*1	<0.01	306	0.53	128	80
JR Sendai Hospital	JR1	Male	48	Painless Thyroiditis	1.8*2	<0.005	198	0.56	184	102*3
	JR2	Male	48	Graves' Disease	5.6*2	<0.005	494	0.56	106	74*3
	JR3	Female	35	Graves' Disease	5.35*2	<0.005	468	0.33	162	85*3
	JR4	Female	42	Thyrotoxicosis*4	2.2*2	0.008	135	0.44	171	92*3
	JR5	Female	49	Thyrotoxicosis*4	1.99*2	0.092	122	0.51	132	80*3
	JR6	Male	51	Graves' Disease	3.28*2	<0.005	391	0.67	165	85*3

*1: The reference range of FT4 in the Tohoku Kosai Hospital Ningen Dock was 0.70-1.48[ng/dL]

*2: The reference range of FT4 in the JR Sendai Hospital Ningen Dock was 0.90-1.70[ng/dL]

*3: Heart rate in JR Sendai Hospital was the pulse rate measured by an automated sphygmomanometer (sitting position), while heart rate in Tohoku Kosai Hospital was determined by ECG (recumbent position).
We corrected the pulse rate to heart rate using a regression formula [14] when calculating predictive probabilities.

*4: No diagnosis was given to patients who did not consult the JR Sendai Hospital Thyroid Outpatient Service.

false positives were then projected onto the SOM and the number of false positives who were located inside the zone of true positives, that is, the number of unavoidable false positives, was counted. We then evaluated the accuracy of the present extended model by comparing the number of our present SOM to that of the SOM in our previous study.

III. RESULTS

A SOM called a grey map that was constructed using a set of 8 parameters is shown in Figure 2A. In the following SOM, true positives (patients with Graves' disease were represented by red numbers, while patients with painless thyroiditis (/thyrotoxicosis without diagnosis) were represented by orange numbers) and healthy individuals (represented by green numbers) were distributed in two distinctive zones. True positives were located on the left, and healthy individuals were located on the right. Patients with severe Graves' disease (high FT4 level) were more likely to be distributed in the deeper area of the left patient zone, whereas those with painless thyroiditis (/thyrotoxicosis who had a mild FT4 level) were more likely to be distributed inside the left patient zone, but closer to the healthy individual zone, such as JR1 (PT), JR4, and JR5. The characteristic distribution in this SOM was similar to our previous findings in which the SOM was constructed using a set of three or four routine tests [13] [14]. These distributional differences made it possible to analyze the characteristics of parameters specific to thyrotoxicosis and use them to differentiate patients.

We showed the component planes (Fig. 2B) accompanying

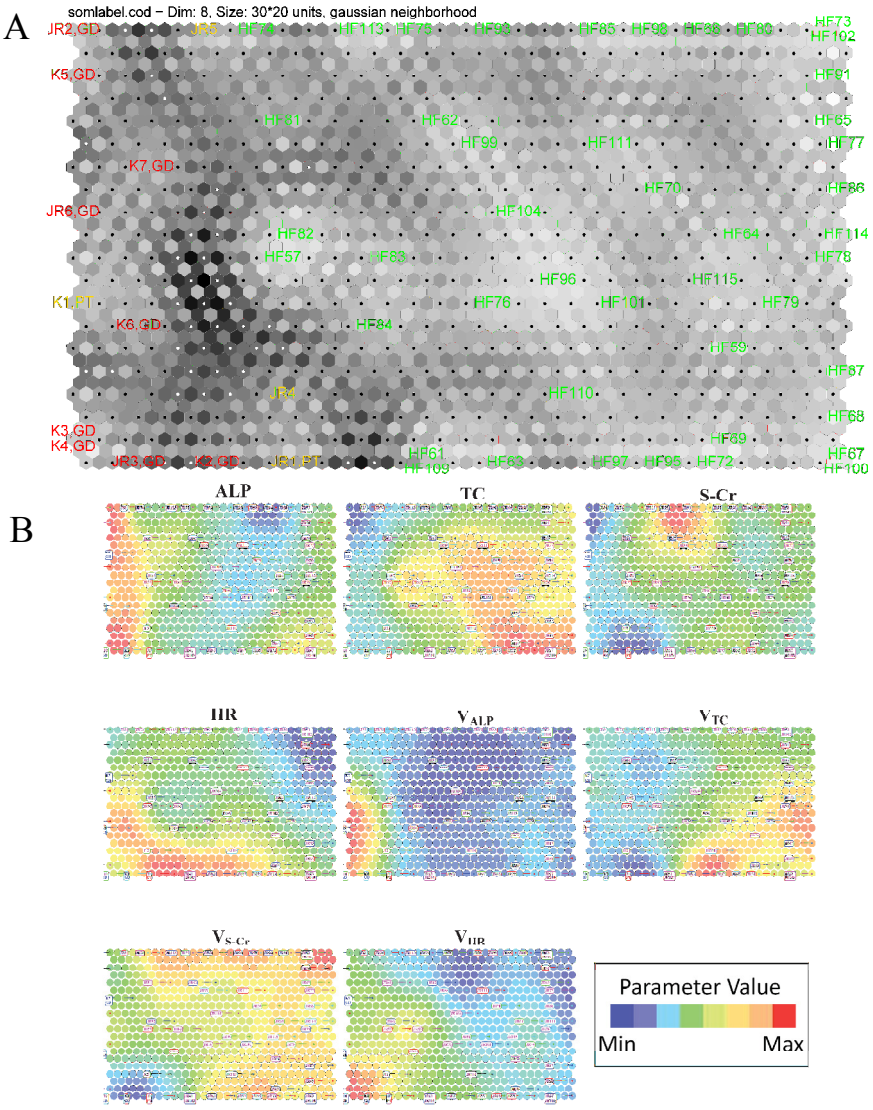


Fig. 2. Features of true positives of thyrotoxicosis were prominently characterized by variations in 8 parameters such as routine tests (ALP, TC, S-Cr, and HR) and their RATV. Red labels show 9 true positives with Graves' disease, orange labels show 4 true positives with painless thyroiditis (/thyrotoxicosis without diagnosis), and green labels show healthy controls.
(A) Grey map of SOM in which true positives and controls were distributed according to the rule that samples that have a similar set of parameters lie nearer, while samples that have different test results lie further away. The calculation was performed under the following conditions: neuron number: 30 × 20, neighboring radius r = 20, learning coefficient = 0.15 and learning frequency: 50,000.
(B) Eight component planes of SOM for each parameter, in which locations of true positives and healthy subjects are all the same as in A, but those with high or low values of these parameters are represented with colors.

the grey map (Fig. 2A), which gave the same location of all samples in Figure 2A, but indicated the level of a given parameter for each sample from high (red) to low (blue). The component planes enabled an instantaneous judgment to identify valuable routine tests in order to discriminate patients with thyrotoxicosis from healthy individuals. The four component planes (ALP, TC, S-Cr, and HR) showed the characteristic ups and downs of the 4 routine tests (slightly elevated ALP and HR and slightly lower TC and S-Cr) in patients with thyrotoxicosis who were distributed over the left area, and another four component planes concerning RATV showed ups and downs in accordance with time-series variations in routine tests. The level distribution of each RATV was almost similar to that of the routine test itself; however, some different distributions were observed in places within the zone. The distribution of v_{HR} was the highest (red) in the lower-left area in which K2 and K4 were located, while the distribution of HR was mild-high (green to yellow) in that area. In contrast, the distribution of HR in the lower-middle area in which healthy individuals HF61 and HF109 were located was the highest (red), whereas that of v_{HR} was mild-low (green) in that area.

Figure 3 shows the same SOM as that shown in Figure 2A; however, 140 false positives (normal of both FT4 and TSH,

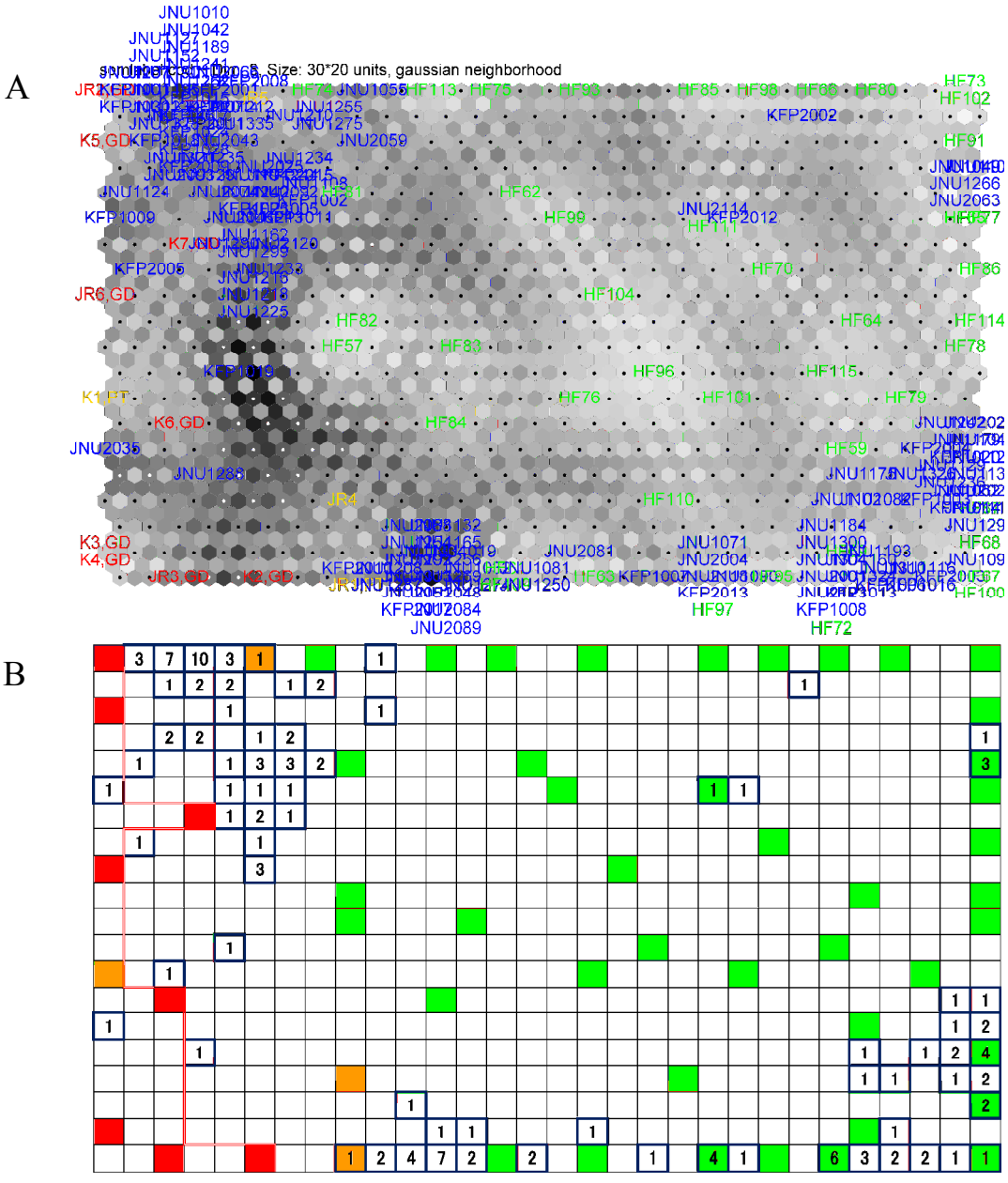


Fig. 3. The same SOM as that shown in Figure 2, but with 140 false positives (with normal FT4 and TSH, represented by blue labels) projected onto it.
(A) Gray map of the original SOM by Kohonen.
(B) A chart-type SOM on the sheet of Microsoft Excel using the same codebook vector obtained in the original SOM shown in 3A.
Each of the red, orange, and green cells denotes that the cell has subjects with Graves' disease, painless thyroiditis (/ thyrotoxicosis without diagnosis), and healthy controls, respectively. Numbers in the blue-bolded cell denote the number of false positives projected onto that cell.

represented by blue labels) were projected onto it. In Figure 3A (grey map of the SOM by Kohonen), approximately half of the subjects were projected around the border region of two groups (thyrotoxicosis (left) and healthy individuals (right)), and only several subjects were located in the deeper left area of the left zone, while the others were dispersedly located in the right healthy region. In order to assist in evaluating these numbers quantitatively, we introduced a chart-type SOM on a sheet of Microsoft Excel, as shown in Figure 3B, using the same codebook vector obtained in the original SOM shown in Figure 3A. Each of the red, orange, and green cells in this chart denoted that the cell was allocated subjects with Graves' disease, painless thyroiditis / thyrotoxicosis without diagnosis, and healthy controls, respectively. Numbers in the blue-bolded cell denoted the number of false positives projected onto the cell. We then drew a threshold line along the cells having patients with Graves' disease who required treatment (indicated by the red line in Fig. 3B), and counted the number of false positives projected on the area surrounded by the line. As a result, the number of unavoidable false positives was significantly reduced from 140 to two while maintaining the true positives of Graves' diseases.

IV. DISCUSSIONS

In the present study, we successfully confirmed the usefulness of considering time-series variations more non-linearly with the SOM by using a set of 8 parameters that included four RATV in addition to the routine tests ALP, S-Cr, TC, and HR. Two groups, patients with thyrotoxicosis and healthy individuals, were separately distributed. Such separated distribution was found in the SOM constructed using a set of three (/four) routine tests, as we previously reported [11]-[14]. It should be noted that a complementary up and down distribution was found among the component planes of each routine test and its RATV. Regarding such a complemental distribution between HR and v_{HR}, most of the patients with Graves' disease had a higher HR, whereas those with a lower HR had a higher v_{HR} .This result suggested that using RATV together with the routine test itself helps to prevent false negatives. In contrast, most of the healthy individuals had normal-lower HR, whereas those with a higher HR had a lower v_{HR} . This result suggested that using RATV together with the routine test itself also helps to prevent false positives. These results may be interpreted as originating from inter-individual variations in heart rate [22]. For example, someone with an athlete's heart will have a lower personal normal range for HR than others; therefore, when they develop Graves' disease, elevations in HR are masked in its absolute value, whereas a relative difference between before and after its onset is expected. In our previous study, the addition of HR to a set of three routine tests yielded many advantages, but an unignorably disadvantage; the HR-added model sometimes avoided false positives and false negatives of Graves' disease, but sometimes caused false positives and false negatives [17]. This limitation was successfully avoided in our present model by considering the RATV of HR in addition to HR itself. We expect these effects in the other routine laboratory tests. Furthermore, by comparing the number of unavoidable false positives in the present study with our previous findings [18], we further decreased this number, which suggests that fully considering the non-linear effect of RATV by adopting a pattern recognition method such as the SOM may be more effective.

Individual differences are known to exist in routine tests, thereby making statistical analyses insufficient. However, we considered the influence of individual differences to be markedly reduced by the introduction of RATV. It became possible to markedly reduce the number of false positives while maintaining that of true positives of Graves' diseases. It is considered easy to track these time-series variations in examinees who visit the Ningen Dock because approximately 70% of visitors are repeaters [23], [24] and missing values are rare; thus, our time-series analysis in the screening demonstrated in this study may be realistic and useful. However, the amount of training data obtained to date is small, and further studies are needed in order to confirm the robustness and versatility of this model by increasing samples hereafter. In addition, it is desirable to perform similar analyses using another PRM such as BRNN and SVM in future research.

REFERENCES

[1] G.J. Canaris, N.R. Manowitz, G. Mayor, E.C. Ridgway: The Colorado thyroid disease prevalence study. Arch Intern Med, 160, pp. 526-534, 2000. http://dx.doi.org/0.1001/archinte.160.4.526.

[2] S.J. Landers: Overlooked, underdiagnosed? Thyroid disease poses a challenge. American Medical News, 47, pp. 25-26, 2004.

[3] S. Reza, A. Shaukat, T.M. Arain, Q.S. Riaz, M. Mahmud: Expression of osteopontin in patients with thyroid dysfunction. PLoS One, 8, e56533, 2013. http://dx.doi.org/10.1371/journal.pone.0056533

[4] M.C. Mosher: Amiodarone-induced hypothyroidism and other adverse effects. Dimens Crit Care Nurs, 30, pp. 87-93, 2011. http://dx.doi.org/10.1097/DCC.0b013e3182052130

[5] M. Beniko, S. Sato, K. Iida, H. Ikawa, H. Kudo, A. Matsumoto, M. Sekiguchi, A. Imamura, Y. Mashio, K. Kawasaki:Evaluation of serum TSH levels in Ningen Dock,Ningen Dock, 24, pp. 885-890, 2009 (In Japanese). http://dx.doi.org/10.11320/ningendock.24.885

[6] K. Kasagi, N. Takahashi, G. Inoue, T. Honda, Y. Kawachi, Y. Izumi: Thyroid function in Japanese adults as assessed by a general health checkup system in relation with thyroid-related antibodies and other clinical parameters. Thyroid, 19, pp. 937-944, 2009. http://dx.doi.org/10.1089/thy.2009.0205.

[7] P.W. Ladenson, P.A. Singer, K.B. Ain, N. Bagchi, S.T. Bigos, E.G. Levy, S.A. Smith, G.H. Daniels: American Thyroid Association guidelines for detection of thyroid dysfunction. Arch Intern Med, 160, pp. 1573-1575, 2000. http://dx.doi.org/10.1001/archinte.160.11.1573

[8] T. Kohonen: Self-organizing maps 3rd edn. Springer-Verlag, Berlin, 2000.

[9] D.J.C. Mackay: A practical Bayesian framework for backpropagation networks. Neural Computation, 4, pp. 448-472, 1992. http://dx.doi.org/10.1162/neco.1992.4.3.448

[10] V.N. Vapnik: The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.
[11] K. Hoshi, J. Kawakami, W. Sato, K. Sato, A. Sugawara, Y. Saito, K. Yoshida: Assisting the Diagnosis of Thyroid Diseases with Bayesian-Type and SOM-Type Neural Networks Making Use of Routine Test Data. Chem Pharm Bull, 54, pp. 1162-1169, 2006. http://doi.org/10.1248/cpb.54.1162
[12] W. Sato, K. Hoshi, J. Kawakami, K. Sato, A. Sugawara, Y. Saito, K. Yoshida: Assisting the diagnosis of Graves' hyperthyroidism with Bayesian-type and SOMtype neural networks by making use of a set of three routine tests and their correlation with free T4. Biomed Pharmacother, 64, pp. 7-15, 2010. http://dx.doi.org/10.1016/j.biopha.2009.02.007
[13] S. Aoki, K. Hoshi, J. Kawakami, K. Sato, K. Satoh, K. Mori, A. Sugawara, Y. Saito, K. Yoshida: Assisting the diagnosis of Graves' hyperthyroidism with pattern recognition methods and a set of three routine tests parameters, and their correlations with free T4 levels: Extension to male patients. Biomed Pharmacother, 65, pp. 95-104, 2011. http://dx.doi.org/10.1016/j.biopha.2010.10.005
[14] S. Aoki, K. Sato, K. Hoshi, J. Kawakami, S. Suzuki, K. Mori, A. Sugawara, Y. Saito, K. Yoshida: improvement of New Low-cost Method for Detecting Abnormal Thyroid Function Making Use of Set of Routine Tests -Addition of Heart Rate, Influence of Dosing and Time Series Drifting are Very Effective-. Ningen Dock, 27, pp. 87-96, 2012. http://doi.org/10.11320/ningendock.27.87
[15] S. Aoki, K. Hoshi, J. Kawakami, K. Sato, W. Sato, K. Satoh, K. Mori, A. Sugawara, Y. Nakagawa, K. Yoshida: Assisting the diagnosis of overt hypothyroidism with pattern recognition methods making use of a set of routine tests, and their multiple correlation with total T4. Biomed Pharmacother, 66, pp. 195-205, 2012. http://doi.org/10.1016/j.biopha.2011.11.018
[16] S. Aoki, K. Satoh, K. Hoshi, J. Kawakami, K. Sato, Y. Saito, K. Mori, K. Yoshida: New Method for Aiding Detection of Abnormal Thyroid Function in Patients Making Use of Set of Routine Tests- Applying It in Screening During General Health Check-ups. Ningen Dock, 26, pp. 9-16, 2011. http://doi.org/10.11320/ningendock.26.9
[17] S. Aoki, K. Sato, K. Hoshi, J. Kawakami, K. Mori, Y. Nakagawa, W. Hida, K. Yoshida: New Low-cost Method for Detecting Abnormal Thyroid Function in Patients Making Use of a Set of Routine Tests-Testing Many More Ningen Dock Examinees and Studying Appropriate Threshold Levels. Ningen Dock International, 2, pp. 19-26, 2014.
[18] S. Aoki, S.Nishizaka, K. Sato, K. Hoshi, J. Kawakami, K. Mori, Y. Nakagawa, W. Hida, K. Yoshida: New clinical decision support system using a set of routine tests to detect thyrotoxicosis in patients: Adding the average rate of annual time-series variations improves diagnostic accuracy. to be submitted.
[19] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen: SOM_PAK: The Self-Organizing Map Program Package. http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml
[20] R.M. Neal: Software for Flexible Bayesian Modeling and Markov Chain Sampling. http://www.cs.toronto.edu/~radford/
[21] C.C. Chang, C.J. Lin: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2, pp. 1-27, 2011. http://dx.doi.org/10.1145/1961189.1961199
[22] H. Kobayashi: Inter and intra individual variations in Heart Rate Variability of Japanese males. J Physiol Anthropol, 26, pp. 173-177, 2007. http://doi.org/10.2114/jpa2.26.173
[23] Committee for Statistical Survey of Ningen Dock Examination: Report on National Aggregate Results of Ningen Dock in 2012. Ningen Dock, 28, pp. 678-690, 2013(In Japanese). http://doi.org/10.11320/ningendock.28.678
[24] Committee for Statistical Survey of Ningen Dock Examination: Report on National Aggregate Results of Ningen Dock in 2013. Ningen Dock, 29, pp. 623-635, 2014(In Japanese). http://doi.org/10.11320/ningendock.29.623

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 11–15
DOI: 10.15439/2015F381
ACSIS, Vol. 6

Queries for Detailed Information System Selection

Agnieszka Dardzinska
Bialystok University of Technology
Department on Mechanics and Computer Science
ul. Wiejska 45c, 15-351 Bialystok, Poland
Email: a.dardzinska@pb.edu.pl

Anna Romaniuk
Bialystok University of Technology
Department on Mechanics and Computer Science
ul. Wiejska 45c, 15-351 Bialystok, Poland
Email:a.romaniuk@doktoranci.pb.edu.pl

***Abstract*—In this paper we assume there is a group of connected information systems forming distributed information system (DS). They work under the same ontology. At the same time, each information system has its own knowledge base. Values of attributes in each information system S form atomic expressions of a language used for communication with others. Collaboration among them is initiated when one of information system S is asked by user to resolve a query containing some nonlocal attributes for S. Therefore it has to contact other information systems to obtain additional, helpful knowledge for finding finally objects satisfying given query. Because there is a set of different information systems connected with a given one, we have to decide which of them is the closest with its knowledge, and which one should be selected by user, for further investigation.**

I. Introduction

IN this paper we assume that there is a group of collaborating incomplete information systems, which are connected with a Query Answering System (QAS) and a knowledge base (K), empty at the begining. By incomplete information system we mean an information system, where attribute values are completely unknown or are connected with corresponding weights [4]. From the definition [3], [4] all the weights for each value of attribute has to sum up to 1. The definition of an information system of type λ presented in this work was initially proposed in [11]. The type λ was presented with a purpose to check all the weights assigned to all values of attributes using Chase algorithm [4], [11]. If a weight is lower than minimal threshold value λ, the corresponding attribute value has to be eliminated. All remaining weights assigned to the rest of attribute values are equally adjusted so their sum is again equal 1. In information systems we take into consideration, we force many semantic inconsistencies. It is related to various interpretations of attributes and their values among sites. In some sites the concept *healthy* can be defined and described completely in a different way than in others. Moreover, in some sites can handle hidden or missing values of attribute *healthy* can be interpreted in a different way.

Quite common method for finding set of objects satisfying the given query from QAS is to replace null values using suggested values obtained from some statistical or rule-based methods. Ontologies [1], [2], [7], [8], [9], [14], [15], [16], [17] are widely used as a part of semantical connection between information systems built independently so they can understand and collaborate with each other. In [11], the notion of the rough semantics which is optimal and a method of its construction was presented. The rough semantics can be easily used to model and handle semantic inconsistencies among sites due to interpretations of incomplete values. As the result of collaborations among information systems, a knowledge base of any system fills up continuously and contains knowledge in form of rules extracted from all information systems [6]. As we mentioned earlier, the names of attributes can be the same among different information systems, but at the same time their granularity levels may differ. Therefore the knowledge base has to satisfy certain properties in order to be used by Chase[11]. We will show that while solving a query, it is worth first to use the knowledge extracted from systems which are semantically close to the given information system.

This work was supported by MB/WM/6/2015

II. Incomplete Information System

Working with information systems in real world, most of data are collected and stored in different independent locations, connected with each other. They form distributed information systems. Therefore it is very possible that some attributes can be missing in some systems and can occur in others.
We say that information system $S(A) = (X, A, V)$ is incomplete and of type λ, if $S(A)$ is an incomplete information system sefined by Pawlak in [10] and these conditions hold:

- X is a set of objects, A is a set of attributes, and $V = \cup\{V_a : a \in A\}$ is a set of all values of attributes
- $(\forall x \in X)(\forall a \in A)[a_S(x) \in V_a$ or $a_S = \{(a_i, p_i) : a_i \in V_a \wedge p_i \in \langle 0, 1\rangle \wedge 1 \leq i \leq m\}]$
- $(\forall x \in X)(\forall a \in A)[(a_S(x) = \{(a_i, p_i) : 1 \leq i \leq m\}) \Rightarrow \sum_i p_i = 1]$
- $(\forall x \in X)(\forall a \in A)[(a_S(x) = \{(a_i, p_i) : 1 \leq i \leq m\}) \Rightarrow (p_i \geq \lambda)]$.

Now, let us assume that $(S_i(A), K)$ for any $i \in I$ are i different incomplete information systems of type λ, and the knowledge base K (empty at the beginning). These systems are represented by the same set of attributes A. The meaning and granularity level of values of attributes from A in information systems S_i is also the same. All information systems $(S_i(A), K)$ communicate with each other, setting up distributed information system (DIS). We assume that, if $a \in A_i \cap A_j$, then the granularity levels of attribute a in two independent information systems S_i and S_j may differ but the meaning of this attribute in them remains the same.

The user submits a query q to the information system S. We extract rules which are in form of a set $R = \bigcup L(R_i)$. They can be later used by any Chase algorithm [4], [5], associated with any of the sites of DIS. Set $L(R_i)$ contains all rules extracted from system S_i. If we want Chase algorithm to be applicable to S, it has to be strictly connected with rules from R which satisfy the following conditions:
- attribute value used in the decision part of a rule from set of rules R has the granularity level either equal to or finer than the granularity level of the corresponding attribute in S.
- attribute used in the decision part of a rule from set of rules R either does not belong to A or is incomplete in S.
- the granularity level of any attribute used in the classification part of a rule from set of rules R is either equal or softer than the granularity level of the corresponding attribute in S.
If we will find a match between the attribute value in information system and the attribute value used in a description of a rule, then two options can be further taken into consideration:
- attribute a involved in matching is the classification attribute in rule. If two attribute values have different granularity level, then the value of attribute a has to be replaced by a finer value (its granularity has to match the granularity of a in S).
- attribute d involved in matching is the decision attribute in rule. If two attribute values have different granularity level, then the decision value d has to be replaced by a softer value (its granularity has to match the granularity of d in S.
In this paper we also take into consideration two information systems and assume that:
Information system S_1 can be transformed into S_2 by containment mapping Ψ ($\Psi(S_1) = S_2$) if following conditions hold:

- $(\forall x \in X)(\forall a \in A)[card(a_{S_1}(x)) \geq card(a_{S_2}(x))]$
- $(\forall x \in X)(\forall a \in A)[card(a_{S_1}(x)) = card(a_{S_2}(x)) \Rightarrow [\sum_{i \neq j} |p_{2i} - p_{2j}| > \sum_{i \neq j} |p_{1i} - p_{1j}|]]$

So, if containment mapping Ψ converts an information system S_1 to S_2, then we can say S_2 is more complete than S_1. It means that for a minimum one pair $(a, x) \in A \times X$, either function Ψ has to decrease the number of attribute values in $a_{S_1}(x)$ or the average difference between any two confidences assigned to each attribute value in $a_{S_2}(x)$ has to be increased by this function.

Let us take two information systems S_1, S_2 both of the type λ, represented as Table 1 and Table 2.

It can be easily noticed that some values assigned to objects in both information systems S_1 and S_2 are different. For example values $a(x_3), a(x_5), a(x_8), b(x_2), c(x_2), c(x_7), e(x_1), e(x_4)$. In each of these cases, an attribute value assigned to an object in S_2 is less general than the value assigned to the same object in S_1. It means that Ψ is a correct containment mapping ($\Psi(S_1) = S_2$).

III. Query processing based on chase and collaboration between systems

Assume that user submits a query $q(B)$ to one of the information system (S, K), which cannot be answered directly as some of the attributes used in a given query are unknown in the primary information system. Assume, we have a group of collaborating information systems working under the same ontology, and user resubmits a query $q(B)$ to other information systems $(S(A), K)$, where $S(A) = (X, A, V)$, $K = \emptyset$, B are the attributes used in $q(B)$, and $A \cap B \neq \emptyset$. Since $(S(A), K)$ can collaborate with itself and other information systems, definitions of hidden or missed attributes for $(S(A), K)$ can be extracted from this information system or from other information systems. It was shown [10] that $(S(A), K)$ can answer the query $q(B)$ assuming that definitions of all values of foreign attributes can be extracted locally or globally and used to finding answer for $q(B)$. Assume now that we have three collaborating information systems with knowledge bases connected with them: (S, K), (S_1, K_1), (S_2, K_2), where $S = (X, A, V)$, $S_1 = (X_1, A_1, V_1)$, $S_2 = (X_2, A_2, V_2)$, and knowledge bases are initially empty $K = K_1 = K_2 = \emptyset$. If the consensus between (S, K) and (S_1, K_1) on the knowledge extracted from $S(A \cap A_1)$ and $S_1(A \cap A_1)$ is closer than the consensus between (S, K) and (S_2, K_2) on the knowledge extracted from $S_2(A \cap A_2)$ and $S_2(A \cap A_2)$, then (S, K) chooses (S_1, K_1) as the detailed information system, more helpful in solving user's queries. All the rules defining unknown attribute values for S are then extracted at S_1 and placed in corresponding K. The problem here is with the predicted values: are they really correctly extracted, and if not, how close they are to the correct values? How can we test this? The classical approach works as follows. First we start with a complete information system, from which we can extract knowledge in form of rules. Then we remove randomly some percent of its values and try to resolve new information system using one of the imputation algorithms. Next we make comparison of the descriptions of objects in both systems: the original one and the system which is the outcome of the imputation algorithm.

Before we will go deeper in the analyzing, we have to make the interpretation of two main functors and and or, denoted in this paper by $*$ and $+$ correspondingly. We can use the semantics of terms proposed in [12] since it preserves distributive property:
For any queries t_1, t_2, t_3 we have $t_1 * (t_2 + t_3) = (t_1 * t_2) + (t_1 * t_3)$. If we consider that $S = (X, A, V)$ is an information system of type λ and t is a term constructed in a standard way from constant values of attributes in V and two functors " $*$ " and " $+$ ", then by the standard interpretation of a term t in S we mean $N_S(t)$ defined as [12]:

- $N_S(v) = \{(x, p) : (v, p) \in a(x)\}$ for any $v \in V_a$
- $N_S(t_1 * t_2) = N_S(t_1) \otimes N_S(t_2)$
- $N_S(t_1 + t_2) = N_S(t_1) \oplus N_S(t_2)$

where

- $N_S(t_1) \otimes N_S(t_2) = \{(x_i, p_i \cdot q_i)_{i \in I \cap J}\}$
- $N_S(t_1) \oplus N_S(t_2) = \{(x_i, p_i)_{i \in I \setminus J}\} \cup \{(x_j, q_j)_{i \in J \setminus I}\} \cup \{(x_i, max(p_i, q_i))_{i \in I \cap J}\}$

Assume that (S, K) is an information system, where $S = (X, A, V)$ and K contains definitions of attribute values in

TABLE I
INFORMATION SYSTEM S_1

X	a	b	c	d	e
x_1	$\{(a_1, \frac{2}{3}), (a_2, \frac{1}{3})\}$	$\{(b_1, \frac{1}{4}), (b_2, \frac{3}{4})\}$	c_1	d_1	$\{(e_1, \frac{1}{2}), (e_2, \frac{1}{2})\}$
x_2	$\{(a_2, \frac{1}{4}), (a_3, \frac{3}{4})\}$	$\{(b_1, \frac{1}{3}), (b_2, \frac{2}{3})\}$		d_1	e_1
x_3		$b_2\}$	$\{(c_1, \frac{1}{2}), (c_3, \frac{1}{2})\}$	d_2	e_2
x_4	a_3		c_2	d_1	$\{(e_1, \frac{2}{3}), (e_2, \frac{1}{3})\}$
x_5	$\{(a_1, \frac{2}{3}), (a_2, \frac{1}{3})\}$	b_1	c_2		e_1
x_6	a_2	b_2	c_3	d_2	$\{(e_2, \frac{1}{3}), (e_3, \frac{2}{3})\}$
x_7	a_2	$\{(b_1, \frac{1}{4}), (b_2, \frac{3}{4})\}$	$\{(c_1, \frac{1}{3}), (c_2, \frac{2}{3})\}$	d_2	e_2
x_8		b_1	c_2	d_1	e_3

TABLE II
INFORMATION SYSTEM S_2

X	a	b	c	d	e
x_1	$\{(a_1, \frac{2}{3}), (a_2, \frac{1}{3})\}$	$\{(b_1, \frac{1}{4}), (b_2, \frac{3}{4})\}$	c_1	d_1	$\{(e_1, \frac{1}{3}), (e_2, \frac{2}{3})\}$
x_2	$\{(a_2, \frac{1}{4}), (a_3, \frac{3}{4})\}$	b_1	$\{(c_1, \frac{1}{4}), (c_2, \frac{3}{4})\}$	d_1	e_1
x_3	a_1	$b_2\}$	$\{(c_1, \frac{1}{2}), (c_3, \frac{1}{2})\}$	d_2	e_2
x_4	a_3		c_2	d_1	e_2
x_5	$\{(a_1, \frac{3}{4}), (a_2, \frac{1}{4})\}$	b_1	c_2		e_1
x_6	a_2	b_2	c_3	d_2	$\{(e_2, \frac{1}{3}), (e_3, \frac{2}{3})\}$
x_7	a_2	$\{(b_1, \frac{1}{4}), (b_2, \frac{3}{4})\}$	c_1	d_2	e_2
x_8	$\{(a_1, \frac{2}{3}), (a_2, \frac{1}{3})\}$	b_1	c_2	d_1	e_3

B. Clearly $A \cap B = \emptyset$. The null value imputation algorithm Chase converts information system $S(A \cup B)$ of type λ to a new more complete information system $Chase(S(A \cup B))$ of the same type. The proposed strategy is new in comparison to known strategies for chasing missing values in relational tables because of the assumption about partial incompleteness of data (sets of weighted attribute values can be assigned to an object as its value). We use ERID algorithm [4] in Chase method to extract rules from such type of data.

IV. SEARCHING MORE DETAILED INFORMATION SYSTEM

Assume again the information system (S, K). As we already mentioned, the knowledge base K, contains rules extracted earlier either locally (inside primary information system) or globally (obtained from distributed information systems). We want to find the most optimal information system (S_i, K) for primary information system (S, K), where by optimal we mean the closest system. The distance between two information systems is calculated using the formula:
$d(S_1, S_2) = \frac{\Sigma_r d_r(S_1 \to S_2) + \Sigma_r d_r(S_2 \to S_1)}{\Sigma_r supS_1 \cdot confS_1 + \Sigma_r supS_2 \cdot confS_2}$
where
$d_r(S_1 \to S_2) = |\frac{supS_2 \cdot confS_2}{max(supS_1,1)} - \frac{supS_1 \cdot confS_1}{max(supS_2,1)}|$.

Information system with the minimal value of $d(S_1 \to S_2)$ has to be chosen. Then it means that S_1 system is the closest to primary information system.

Let us assume we have distributed information system consisting of three different information systems, collaborating with each other S, S_1, S_2, represented as Table3, Table4 and Table5, respectively.

TABLE III
INFORMATION SYSTEM S

X	a	b	c	d
x_1	2	2	$-$	
x_2	2	1	$+$	
x_3	0	1	$-$	
x_4	1	2	$+$	
x_5	0	1	$-$	
x_6	1	3	$-$	

TABLE IV
INFORMATION SYSTEM S_1

Y	a	b	c	d	e
y_1	1	1	$+$	1	L
y_2	1	2	$-$	1	H
y_3	2	3	$-$	1	L
y_4	2	1	$+$	1	L
y_5	1	2	$-$	0	H
y_6	2	3	$-$	0	H
y_7	0	1	$+$	1	L

Information system S received a query $q(B) = (a, 0) * (b, 1) * (d, 1)$ and it has no information about attribute d, which is hidden. Meanwhile this attribute appears in other systems. The purpose is to choose one of the systems: either

TABLE V
INFORMATION SYSTEM S_2

Z	a	b	c	d
z_1	2	1	+	0
z_2	2	1	+	1
z_3	1	2	−	0
z_4	1	3	−	1
z_5	0	2	−	0
z_6	0	3	+	1

S_1 or S_2, from which values of attribute d in system S can be predicted. After this step we will be able to find set of objects satisfying given query $q(B)$. Because attributes a, b, c appear in all the systems, first we extract, from each information system independently, rules describing a, b, c in terms of other attributes, using method similar to LERS. Next, for each rule we calculate support and confidence in a standard way [4], [5].

For system S_1 we have:
$(b,1) \rightarrow (a,1)$ with $sup = 1$, $conf = \frac{1}{3}$
$(b,1) \rightarrow (a,2)$ with $sup = 1$, $conf = \frac{1}{3}$
$(b,1) \rightarrow (a,0)$ with $sup = 1$, $conf = \frac{1}{3}$
$(b,2) \rightarrow (a,1)$ with $sup = 2$, $conf = 1$
$(b,3) \rightarrow (a,2)$ with $sup = 2$, $conf = 1$
$(c,+) \rightarrow (a,1)$ with $sup = 1$, $conf = \frac{1}{3}$
$(c,+) \rightarrow (a,2)$ with $sup = 1$, $conf = \frac{1}{3}$
$(c,+) \rightarrow (a,0)$ with $sup = 1$, $conf = \frac{1}{3}$
$(a,1) \rightarrow (b,1)$ with $sup = 1$, $conf = \frac{1}{3}$
...
$(a,1) * (c,+) \rightarrow (b,1)$ with $sup = 1$, $conf = 1$
$(a,1) * (c,-) \rightarrow (b,2)$ with $sup = 2$, $conf = 1$
...
For system S_2 we have:
$(b,1) \rightarrow (a,1)$ with $sup = 2$, $conf = 1$
$(b,1) \rightarrow (a,2)$ with $sup = 2$, $conf = 1$
$(b,2) \rightarrow (a,1)$ with $sup = 2$, $conf = \frac{1}{2}$
$(b,2) \rightarrow (a,0)$ with $sup = 2$, $conf = \frac{1}{2}$
$(b,3) \rightarrow (a,1)$ with $sup = 2$, $conf = \frac{1}{2}$
$(b,3) \rightarrow (a,0)$ with $sup = 2$, $conf = \frac{1}{2}$
$(c,+) \rightarrow (a,2)$ with $sup = 2$, $conf = \frac{2}{3}$
$(c,+) \rightarrow (a,0)$ with $sup = 1$, $conf = \frac{1}{3}$
$(c,-) \rightarrow (a,1)$ with $sup = 2$, $conf = \frac{2}{3}$
$(c,-) \rightarrow (a,0)$ with $sup = 1$, $conf = \frac{1}{3}$
$(a,1) \rightarrow (b,2)$ with $sup = 1$, $conf = \frac{1}{2}$
...
$(a,0) * (b,3) \rightarrow (c,+)$ with $sup = 1$, $conf = 1$
$(a,0) * (d,1) \rightarrow (c,-)$ with $sup = 1$, $conf = 1$
...
We do the similar procedure for system S.
Next the distance between S and S_1 is calculated: $d(S \rightarrow S_1) = \frac{33.36+20.3130+34.66}{36,66+45,3130+39,45} = 0.686$ and so between S and S_2: $d(S \rightarrow S_2) = \frac{33.36+17.17}{30+29.66} = 0.85$ Because the distance between S and S_1 is smaller than between S and $_2$ (the factor is smaller), we choose S_1 as more detailed information system for contact with S. From chosen information system S_1, rules describing attribute d in terms of a, b, c are extracted. We can use algorithm ERID [4], [11], [12] for extracting rules from incomplete information system and put them into knowledge base K. These rules also allow us to uncover some hidden attribute values in information system S. Therefore the submitted query $q(B)$ can be answered and the set of objects is x_3, x_5.

This method can enhance and expand the scope of decision support system, which assists patients and physicians with the challenge of managing pancreas diseases.It involves data collection of pancreatic cancer, risk factors, common characteristics and survival rates. The objective of the system is to detect problems in pancreatic management and to recommend changes to correct these detected problems. It is very important problem, as the survival rate of pancreatic cancer is very poor and those that survive are due to early detection. Early detection of pancreatic cancer is very difficult and often involves some sort of invasive testing procedure. One of the main symptoms of pancreatic cancer is chronic pancreatitis. If physicians are prompted to order additional testing at the time they entered a pancreatitis diagnosis the cancer has a chance to be detected earlier.

V. FINAL REMARKS

In the paper we proposed the method of finding the closest information system to the client. Our goal was to find the best information system, which will be helpful in answering the query submitted to the client. Once we find more detailed information system to the given one, we are able to build knowledge base consisting of rules extracted in distributed systems. Then we can apply the rules, so changes of values of attributes in a query $q(B)$ can be made. The unknown attributes can be replaced by mixture of attributes which are present in both of the systems, the query will transform into more clear and understandable form and can be then answered.

This method was initially tested in medical databases with special preferences related to different pancreas diseases and gives some promising results. Several information systems, keeping different information were connected with other in hospital systems. Each information system consists of a big set of knowledge not necessary connected with patients with pancreas diseases. From the whole set of databases - systems with the best knowledge about particular diseases should be chosen for communication. The best treatment for most of pancreatic diseases, especially cancer, depends on how far it has spread, or its stage. The stages of pancreatic cancer are quite easy to understand and work on. The problem which appears in pancreas cancer treatment is how to describe the stage of pancreatic cancer without previously resorting to major surgery and how to minimize the risk of the disease. In practice, doctors choose pancreatic cancer treatments based upon imaging studies, surgical findings, and an individual's general state of well being. Determining pancreatic cancer's stage is often not so easy. Imaging tests like CT scans and ultrasound of course give doctors some information, but

knowing exactly how far pancreatic cancer has spread already, usually requires deep surgery. Therefore finding the closest information system help to work with the latest, most advanced therapies for pancreatic diseases and can be used to ensure the most advanced treatment with the least impact on patient's body. Using our method we can suggest some restrictions showing how to reduce the risk of pancreatic disease. We obtained knowledge which gives information that smoking is the most important avoidable risk factor for pancreatic cancer, and quitting smoking helps to lower risk. Also getting physical activity and eating well can help patient to stay at a healthy weight and also reduce the risk of illness. The third possibility to minimize the risk for pancreatic cancer is to avoid workplace exposure to harmful substances such as certain pesticides and other chemicals.

REFERENCES

[1] V.R. Benjamins, D. Fensel, A.G. Prez, "Knowledge management through ontologies", *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management* (PAKM-98), Basel, Switzerland, 1998

[2] B. Chandrasekaran, J.R. Josephson,V.R. Benjamins, "The ontology of tasks and methods", *Proceedings of the 11th Workshop on Knowledge Acquisition, Modeling and Management*, Alberta, Canada, 1998

[3] A.Dardzinska, "Action rules mining", *Springer Verlag*, 2013

[4] A. Dardzinska, Z. Ras, "Chase2, Rule based chase algorithm for information systems of type lambda", *Proceedings of the Second International Workshop on Active Mining*, Maebashi City, Japan, LNAI SpringerVerlag no3430, 2005

[5] A. Dardzinska, Z. Ras, "On Rules Discovery from Incomplete Information Systems", *Proceedings of ICDM'03 Workshop on Foundations and New Directions of Data Mining*,Melbourne, Florida, IEEE Computer Society, 2003

[6] A.Dardzinska, A.Romaniuk, "ncomplete distributed information systems optimization based on queries", *Advances in Swarm and Computational Intelligence : 6th International Conference : ICSI 2015*, LNCS Springer, Pekin, China, 2015

[7] D. Fensel, "Ontologies: a silver bullet for knowledge management and electronic commerce", Springer-Verlag, 1998

[8] N. Guarino, "Formal Ontology in Information Systems", IOS Press, Amsterdam, 1998

[9] N. Guarino, P. Giaretta, "Ontologies and knowledge bases, towards a terminological clarification", *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, IOS Press, 1995

[10] Z. Pawlak, "Information systems - theoretical foundations", *Information Systems Journal*, Elsevier, Vol. 6, 1981

[11] Z. Ras, A. Dardzinska, "Ontology based distributed autonomous knowledge systems", *Information Systems International Journal*, Elsevier, Vol.29, No. 1, 2004

[12] Z. Ras, A. Dardzinska, "Solving Failing Queries through Cooperation and Collaboration", *World Wide Web Journal*, Springer, Vol. 9, No. 2, 2006

[13] Z. Ras, S. Joshi, "Query approximate answering system for an incomplete DKBS", *Fundamenta Informaticae Journal*, IOS Press, Vol. 30, No. 3/4, 1997

[14] J. Sowa, "Ontology, metadata, and semiotics", *Conceptual Structures: Logical, Linguistic, and Computational Issues*, LNAI, No. 1867, Springer-Verlag, Berlin, 2000

[15] J. Sowa, "Knowledge Representation: Logical, Philosophical and Computational Foundations", *Brooks/Cole Publishing Co.*, Pacic Grove, CA, 2000

[16] J. Sowa, "Ontological categories", *Shapes of Forms: From Gestalt Psychology and Phenomenology to Ontology and Mathematics*, Kluwer Academic Publishers, Dordrecht, 1999

[17] G. Van Heijst, A. Schreiber, B. Wielinga, "Using explicit ontologies in KBS development", *International Journal of Human and Computer Studies*, Vol. 46, No. 2/3, 1997

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 17–24

DOI: 10.15439/2015F416
ACSIS, Vol. 6

Self-Explanation through Semantic Annotation: A Survey

Johannes Fähndrich, Sebastian Ahrndt and Sahin Albayrak
DAI-Labor, Technische Universität Berlin
Faculty of Electrical Engineering and Computer Science
Berlin, Germany
Email: johannes.faehndrich@dai-labor.de

***Abstract*—Semantic information is considered as foundation upon which modern approaches attempt to tackle the challenges of dynamic environments – service orchestration and ontology matching are two examples for the use of such information. Yet, many developers avoid the additional effort of adding semantic information (e.g., through annotations) to their data sets – limiting the reusability and interoperability of their Apps, services, or data. This problem is called the "knowledge acquisition bottleneck", which can be addressed by providing suitable tool support. This survey analyses the state-of-the-art of such tools that support developers in the task of semantically enriching entities. Providing an overview of available tools from the early days until now, we particularly focus on the 'level of automation'. Concluding that automation is very limited in contemporary tools we propose a concept that mixes connectionist and symbolic representation of meaning to decrease the manual effort.**

I. Introduction

'ONE of the most complex construction tasks humans undertake' [1, p. 1] is the development of distributed software systems that are intended to solve complex real-world tasks. Such systems, which become ever more interconnected and diverse, evolve over time. One can imagine, that this leads to heterogeneity problems, as different parties at different times make use of different technologies to reach their goals. Describing the involved entities (devices, services) of such systems in an structured and machine readable way is still an unsolved research issue and a human driven task. As the creation of these descriptions is only partially feasible at design time, many developers avoid adding semantic information to their data sets in order to save additional efforts – thus neglecting the advantages of an enrichment with additional (semantic and contextual) information. This problem is called the "knowledge acquisition bottleneck" [2].

To ease the semantic enrichment many approaches and tools are available, each one able to support developers in the task of adding semantic information to critical data. This work surveys those tools. Following the trend of leaving more and more issues dealt at runtime, we want to put a particular focus on the 'level of automation' that is provided by the available approaches. Hence, this work gives an overview on methods, tools and approaches, that particularly concentrates on the amount of automatism provided and the possibilities for the user to interact with the annotation process. Additionally, we focus our attention on the self-explanation property, interpreted as self-explanatory descriptions of software components as they occur in service oriented architectures or agent oriented architectures. The tools surveyed are used in creating such self-explanatory descriptions for artificial reasoners that use them during run time to couple distributed systems.

In order to identify relevant candidates, we carried out a literature research using the following databases, sources, and keywords:

- Search engines: Google Scholar[1], ACM Digital Library[2], IEEE Xplore Digital Library[3], JSTOR[4], Papers3[5], Springer Link[6]
- Proceedings crawled: International Semantic Web Conference Series (ISWC), International Joint Conference on Artificial Intelligence Series (IJCAI), AAAI Conference on Artificial Intelligence Series (AAAI), International World Wide Web Conference Series (ACM WWW), The Journal of Web Semantics, Conference on Hypertext and Social Media Series (ACM HYPERTEXT), Human-Computer Interaction Series (CHI)
- Keywords used: Semantic annotation tools, (semi-) automatic semantic annotation, semantic tagging, (semi-) automatic semantic tagging, ontology annotations, annotation tools for the semantic web

The conferences where crawled from the year 2000 until 2014 if available. Further interesting publications were found by specifically looking into the lists of references.

The remainder of the paper is structured as follows: Next we will introduce existing surveys and compare their results with regards to the specific focus of this work (See Section II). Afterwards, we will present the survey results, starting with the topology we used to classify all considered approaches (See Section III). Subsequently, we discuss the survey results and give some insights into future research challenges. To substantiate the results we also propose how a automatic annotation tool should be structured (See Section IV). Finally, we wrap up with a conclusion (See Section V).

[1]Further information: https:/scholar.google.de/
[2]Further information: http:/dl.acm.org/
[3]Further information: http:/ieeexplore.ieee.org/Xplore/home.jsp
[4]Further information: http:/www.jstor.org/
[5]Further information: http:/www.papersapp.com/mac/
[6]Further information: http:/link.springer.com/

II. Related Work

The development of ontologies has a long history as they were identified timely as practical to conceptualise data [3]. Within the process of semantically annotating data the ontology defines the vocabulary and structure of the annotation result. Starting at this point, we are able to distinguish three generation of tools that support the semantic enrichment:

- The *first generation* are browsers, whose main purpose is viewing the semantic information, which is also called the graph of things [4].
- The *second generation* are annotation tools offering the capability to view and to modify the semantic information.
- The *third generation* are approaches that in addition to viewing and modifying the semantic information offer the option to adapt the underlying ontology.

As these generations are different in nature, the field of surveys with respect to ontology development is it also. For example, *Lopez* [5] and *Braun et al.* [6] present surveys about methodologies of ontology development. In addition, several more recent surveys describe annotation and querying tools (cf. [7], [8], [9], [10]). On the other hand, *Islam et al.* [11] is giving a short but holistic overview over methodologies, standards and tools for the semantic web. Covering the third generation of approaches, *Ding and Schubert* [12], *Gomez-Perez et al.* [13] and *Drumond and Girardi* [14] surveyed ontology learning methods. These works omit the connection to practical applications.

In conclusion, our literature research shows that although there are many surveys available, they rather focuses on methods of information retrieval than to consider the aspects of AI-methods (in particular the level of automation) in practical applications.

III. Tools

The aim of this section is to present an overview of the state-of-the-art regarding tools and approaches used to create and mange semantic information. To ease the reading, we will further refer to an approach, method or tool using the term *solution*. These solutions reach from informal "best practices" like Hash-Tags over Microformats to standards like RDFa (Resource Description Framework in attributes). The critical reader might think of this range of solutions as to broad, since we compare ontology editors like Protege [15] with Browser plugins like Biggy Bank [16]. Since the goal of this survey is to collect solutions and possible extension points to overcome the knowledge acquisition bottleneck, we argue that all mentioned solutions can be used to annotate semantic information to given text — e.g., from Webpages. Here, we want to clarify that it is beyond the scope of this work to judge the language used for annotation, its expressiveness or purpose of use.

In order to classify the examined solutions we utilise different properties. Firstly, as the focus of this survey is based on the degree of automation, we want to classify the presented approaches based on the capability to structure unstructured information in an automated, semiautomatic and/or manual way. Hence, we distinguish the degree of *automation* in four categories:

1) **None**—None means that there is no automatism available. This implies, that all tasks have to be performed by a human.
2) **Semi**—Semi describes the ability to automatically perform some task with the constraint that there is still the requirement to supervise the process.
3) **Collection**—Collection describes the ability to automatically collect information. Since the collection of information is a time consuming task, the exploration of for example deep web annotations [17] can be automated. The extraction of structured information still requires human intervention.
4) **Full**—Full means the capability to collect information, extract additional information (e.g. annotations) as well as integrating new information into the information source without any human intervention. This is not considered restricted by the possibility of a manual annotation, but rather to propagate and integrate the newly gained information.

Besides the automation aspect, we want to clarify whether a solution is *platform independent* or not. This is important to the heterogeneous character of hard- and software used in smart environments (and the semantic web). Here, we also have to take into account the used *language*, which describes the semantic information. As mentioned above, there is the trend to leave more details dealt at runtime. Hence, the examined solutions are also classified to a property called *online*—meaning solutions that are able to integrate new information, enabling users to browse several information sources and collect information during runtime. To prevent the overwhelming of the user with the provided amount of information, we analyse the *search* capability of each solution as well. Furthermore, as some information may be of private manner the descriptive semantic information is it, too. The ability of a solution to decide which information should be shared and which should be kept secret is called *privacy*. One inherent feature here is the ability to share information. We refer to this feature with the term *sharing*. In the end, the used classification takes into account some technical aspects: *Extensible* and *UI* (User Interface). The latter one describes the way a solution represents itself to the user. Since the semantic web community provides most of the solutions surveyed in this work, the UI is mostly a web site beside some exception like frameworks. The first one—extensible—describes the ability to add new functionality to the solution. Choosing a solution that is not able to adapt to new requirement may be fatal for future work on this topic.

The classification of the examined solutions is illustrated in Table I. In order to give some more structure to the results, we proceed with a description of the considered solutions based on their essential functionality and classified according to the three generations mentioned above.

TABLE I: The survey results for all examined solutions.

	Solution	Language	Platform ind.	Online	Search	Privacy	Sharing	Extens.	Automation	UI
1	Disco [18]	RDF	×	×	–	–	–	–	Collection	Web
2	MSpace [19]	RDF	×	–	–	×	×	–	None	Web
3	RDFa Developer [20]	RDFa, Micro	–	×	×	×	–	–	None	–
4	Flamenco [21]	–	×	–	×	–	–	–	None	Web
5	Oink [22]	RDFs	–	–	–	–	–	–	Collection	Web
6	Longwell [23]	RDF	×	–	×	–	–	–	None	Web
7	Sigma [24]	RDF	–	–	×	–	–	–	Collection	Web
8	Aquabrowser [25]	–	–	–	×	–	–	×	None	Web
9	Freebase Parallax [26]	RDF	–	×	–	–	–	–	Collection	Web
10	Tabulator Extension [4]	RDFs, OWL	–	×	–	–	×	×	None	–
11	RangeAnnotator [27]	RDF	–	–	–	×	×	–	None	Web
12	GrOWL [28]	OWL	–	–	×	–	–	–	None	Forms
13	OntoStudio [29]	multiple	–	–	–	–	–	×	Semi	Forms
14	Swoop [30]	OWL	×	–	–	–	–	×	Semi	Web
15	GKB Editor [31]	–	–	–	–	–	–	–	None	Forms
16	Melita [32]	XML	×	×	–	–	–	–	Semi	Forms
17	OBO Edit [33]	OBO	×	–	–	–	–	–	None	Web
18	Magpie [34]	RDF	–	×	–	×	–	×	Collection	Web
19	DOME [35]	RDFS, OWL	×	–	–	×	×	–	Collection	Web
20	Biggy Bank [16]	RDF	–	–	×	×	×	×	None	Web
21	Semantic Turkey [27]	RDF	–	–	–	×	–	×	Semi	Web
22	UIMA Web Anno. [36]	RDF, OWL	×	×	–	×	–	×	Semi	Web
23	Haystack [37]	RDFs	–	×	–	–	×	×	None	Web
24	IBM EODM [38]	RDF, OWL	×	–	–	–	–	×	None	Forms
25	Topia [39]	RDF	×	–	×	–	–	–	Collection	Web
26	Protege [15]	RDF, OWL	×	–	–	–	–	×	None	–
27	Scooner [40]	RDF	–	–	×	×	–	–	Collection	Web
28	Morla [41]	RDF	–	×	–	–	×	×	None	Forms
29	CmapTools O.E. [42]	RDF, OWL	–	–	–	–	×	×	None	Forms
30	Chimær [33]	RDF, DAML	×	–	–	–	–	–	None	Web
31	KAON2 [43]	OWL-DL	×	–	–	–	–	×	Semi	–
32	Knoodl [44]	RDF, OWL	×	–	×	–	–	–	None	Web
33	Virtual Ontology Modeler [45]	RDFs, OWL	×	–	×	–	×	–	None	Web
34	Ontolingua [46]	KIF	×	–	–	–	×	–	None	Web
35	Moki [47]	OWL	×	–	×	–	×	×	Collection	Web
36	OntogGen [48]	RDFS, OWL	–	–	–	–	×	–	Semi	Forms
37	MnM [49]	KMi	–	×	×	–	×	–	Semi	Web

A. The 1st Generation

We refer to this generation of solutions as browsers. The main purpose of browsers is viewing the graph of things, which is manifested in annotations that are attached to the web of documents. *Berners-Lee et al.* [50] introduce the Tabulator browser. Here, the semantic meta information about some resource is collected and displayed in tables. Tabulator allows the user to search through the presented information and to group them by sources; but not to modify them. Several other solutions for viewing semantic information are available. They can be subsumed under the term semantic browser (e.g., [4], [19], [21], [23], [25], [51], [52], [53], OpenLink Data Explorer[7], Zigist[8], Marbles[9]). In this work we put particular interest on methods on editing the semantic information. Therefore, we can neglect most of the first generation solutions. However, on behalf of the interested reader, we can refer to surveys which focus such solutions (cf. [7], [8], [9], [11]).

As a first step towards a broader function range, *Disco* [18] additionally used the index Sindice[10] to collect semantic information online. Following a similar approach, *Sigma* [24] automates the collection and consolidation from multiple information sources and is focused on collecting and viewing the entities resulting from a query. Although, the main purpose of first generation solutions is viewing the information, this does not mean that there is no automatism. The *Aquabrowser* [25], for example, indexes the information made available to it and creates bags of words and facets without human intervention. Here, the interested reader is pointed to *Stepfaner et al.* [54], which introduce a taxonomy of faceted search. Following a different approach, the *Freebase Parallax* [26] solution can be classified as set-based browser. This type of solution allows switching between properties collected in sets [54]. These approaches are illustrated in Table I at the positions 1–9.

B. The 2nd Generation

The restriction of read only solutions, leads us to the second generation—namely annotation solutions. Second generation solutions have the capability to modify the semantic information. Tools like *GrOWL* [28] and *Knoodl*[44] offer the

[7]For further information, refer to http://ode.openlinksw.com

[8]For further information, refer to http://dataviewer.zitgist.com/

[9]For further information, refer to http://marbles.sourceforge.net/

[10]For further information, refer to http://sindice.com/

capabilities to view and edit semantic information described within ontologies. They can be classified as prototypes of the second generation. Due to the wealth of such solutions, we will further describe only those introducing new functionalities.

Berners-Lee et al. started tiptoeing towards the editing of public semantic information with *Tabulator Redux* [4]. In their work, the discuss where the semantic information should be stored. Ignoring privacy issues, Tabulator Redux enables user to add semantic information to a public wiki. Following *Ciravegna et al.* [32], it seams reasonable that users should create annotations, as they can annotate their points of interest at nearly no cost (on-the-fly). Furthermore, the authors introduced different requirements that must be accomplished and that are tackled by their own solution. *Melita* [32] addresses multiple usability issues arising from the pro-activeness of the user and separates the annotation process into two phases: The training phase, where the user adds annotation manually and the active annotation phase, where the system adds semantic information automatically. During the training phase the user is supported by the learning algorithm $(LP)^2$ [55], which enables an automated annotation behaviour. A similar solution is represented by *Amilcare* [56] also using the $(LP)^2$. An additional feature is introduced by *Magpie* [34]. Magpie allows the use of ontologies to annotate elements of websites. Furthermore, it enables user to specify services associated with the annotation entities. This functionality leads to the automation of ontology development by leaving the architecture open for new services. *Chimæra* [33] describes another aspect of the creation and maintaining of semantic information. The authors argue that ontologies should be created in a distributed manner and propose approaches to maintaining and merging semantic information in ontologies. The *DERI Ontology Management Environment* [35] (DOME) is a specialized ontology editing and maintaining concept, focused on a 'community-driven ontology management'. Hence, the focus lies on alignment, versioning and aggregation. As DOME focuses on the distributed maintaining of semantic information an automatism has been established to populate information. These approaches are illustrated in Table I at the positions 10–20.

C. The 3rd Generation

The second generation solutions allow the user to import existing ontologies for further use. Missing here is the capability to extend these ontologies, which leads us to the third generation of solutions, which are context-aware. Meaning that these solutions are able to adapt the used ontology to the context of use. Here, *Pazienza et al.* [27] introduce the *Semantic Turkey*, an extension of the Firefox browser, which was originally developed as a semantic bookmarking tool in 2007 [57]. In a further development stage, it was combined with the *RangeAnnotator* [27] enabling the extraction of information encoded in RDFa and Microformats. The extracted information are integrated into an *UIMA*[11] process. In addition, the RangeAnnotator adds the capability of *Xpointers*[12]. Another solution based on the Semantic Turkey framework, presented by the same research group is *STIA* [58], an annotation tool to organise pertinence between laws. *Fiorelle et al.* [36] present an additional extension of the Semantic Turkey named *UIMAST Web Annotator*. Here, structured information as in HTML or PDF documents can be annotated and used to enrich user defined ontologies. This process is called Computer Aided Ontology Development (COD). Consequently, their approach is proposed as COD Architecture (CODA) with the goal of semi automatic ontology creation. It is open to extensions during runtime by using the OSGi[13] standard. Following a similar approach, *Scooner* [40] integrates several information extraction techniques to boot strap concepts out of a knowledge base. *OntoGen* [48] extends this automatism using multiple artificial learning approaches that support the user during the creation process by proposing comparable concepts of existing ontologies.

After having created tools to work with ontologies, the semantic web community fostered their technologies to feed back into semantic tools like *Haystack* [37]. Haystack uses RDFa to describe functionalities and user interfaces with the goal to create web applications. The crux of Haystack lies in the orchestration of services producing the functionality in the background and presenting their results to a user. The *CmapTools Ontology Editor* [42] describes the formalisation problem of unstructured information to structured information in a concept map-based manner. Furthermore, they distinguish between expert, experienced and normal users by adapting the user interface to ease the introduction phase to the user.

In contrast to previous solutions, frameworks exist which offer extensive features for the development of ontologies as e.g., the *EMF Ontology Definition Metamodel* [38] (EODM). One can imagine, that there are solutions that can not clearly be marked as frameworks for developers or as development suits for the creation of ontologies without any programming. *Protege* [15] and it counterparts *Ontosautus* [59] and the *Generic Knowledge Base Editor* [60] can be located between both worlds. Another research challenge is addressed by the *Topia* [39] project. Here the use of semantics is discussed within the generation of hypermedia [61]: 'The Topia project is developing a system that generates presentation structure around media objects returned from semantic-based queries.' [39]. Therefore Topia offers capabilities to combine informations from multiple sources concerning one topic using ontology matching techniques. With the *Modeling Wiki (Moki)* [47] a solution for user generated content is presented, which allows to extend a semantic wiki with formal ontologies. These structured descriptions can be interpreted as self-explanatory, depending on the amount of information modelled as formal semantics.

One of the most advanced annotation frameworks is created with *MnM* [49]. After a manual annotation, the MnM

[11] Unstructured Information Management Architecture (UIMA) - http://uima.apache.org/

[12] For further information, refer to http://www.w3.org/TR/xptr-framework/

[13] Open Services Gateway Initiative Framework (OSGi) - http://www.osgi.org/

framework is able to annotate new documents automatically. Although, MnM is based on a rather specific ontology language (KMi) it stores its annotations in a ontology and is able to extract annotations automatically after a learning phase.

It seams that with the advancement in this research, the goal of creating self-explaining elements is getting into reach. These approaches are illustrated in Table I at positions 21–37.

IV. Discussion

In our survey we analysed approaches that allow for the observation and editing of semantic information. Based on this survey we can state that many tools have emerged in the semantic web community. As it was our intention to classify analysed tools by their 'ease of use', we want to put up the respectively identified 'level of automation' for discussion. To start with, none of the examined approaches was able to work in a fully automated fashion. Thus, we can emphasise that there is still a difference between the stated aims of semantic research and the reality. In our opinion, semi-automatisms or solutions that are capable of learning can be considered the bleeding edge. However, any semi-automation involves human interaction, which implies that user interfaces have to be provided [62]. In our opinion, sharing semantic information is another very promising concept, however, this mechanism puts another issue in focus: privacy. Admittedly, privacy is an important issue whenever data is made available, yet, matters of privacy are far beyond this work. We leave such considerations open for future works and endorse the concept of sharing semantic information as a very capable one. Using technology independent standards for the description of semantic information may additionally further the acceptance for this mechanism.

Whenever information sources are updated (either by means of annotations or by automated procedures), the speed at which the updated information become available, plays an important role. If the update occurs (almost) immediately, we refer to the process as being 'online' capable. Online capability allows users to make annotations while browsing data sets. This feature may foster semantic annotation processes to be a natural part of browsing. Furthermore, when it comes to the Internet, finding and retrieving data can be considered as a constituting functionality. However, an ever increasing amount of information makes this task difficult and fosters the use of semantic enrichment of datasets. We therefore argue, that tools have to account for sophisticated search routines. Referring to our main intention, that is, to identify promising tools for further extension, we want to conclude at this point.

Taking the above mentioned properties into account, the general trend in this research area becomes fairly apparent. To foster the (automated) derivation of self-explaining information, approaches such as UIMA Web Annotator (CODA) seam to be worth extending. Admittedly, the current version of CODA is still miles away from the stated aims of semantic research, where 'everybody might say anything about anything' [4], yet, in our opinion, CODA is the most promising approach to achieve this goal.

A. Research Challenges

An AI that should be able extract sense or meaning from texts requires the ability to learn new meaning by itself and, thus, requires the ability to explain new words to itself. We defined this ability in a prior work [63] and within this work substantiated that there are still many hurdles that must be addressed to achieve this objective:

Meaning itself need to be represented in an appropriate way (in a formal manner) to be handled by an AI. Since meaning is not precisely defined, this is subject to research. We will look at meaning in the linguistic sense, which can be defined as follows: Meaning is what the source of an expression (message) wanted the observer to infer from the expression [64]. Since semantics is the theory on how meaning is transferred, a semantic transference and interpretations process is required. There are four parts for the meaning of a word which are of concern to an AI:

- *Denotation*: The so called denotation represents the primary or basic meaning of a word. This can be seen as the definition of a word that is represented in some kind of mental lexicon (or a dictionary).
- *Connotation*: The connotation is the abstract idea presented by the word. This can be seen as the conceptual representation of the meaning of a word. This includes the connectionist interpretation of meaning since here the meaning is interpreted as the unity of its relations to other concepts.
- *Conceptualisation*: To be able to come up with a conceptual representation of the meaning of a word, one needs to abstract from the word to a specific concept (*i.e.*, one needs to connect the word with a known concept). This process is named Conceptualisation and helps to clarify a word within a language.
- *Pragmatics*: The meaning of words is not independent of the context the words are used in. Thus an context dependent representation of meaning (a pragmatic one) has to be created (e.g. mouse (computer) vs. mouse (pet)).

Furthermore starting from the meaning of one word, the meaning of sentences need to be extracted. We neglect this here, since it is seen as a next step after having a meaningful representation of a single word.

Technically adding semantic information generally rises the question on how to make this data available: publicly available or with restricted access. Firstly, semantic information might be directly attached to the respective dataset. However, this implies the source to be editable, which furthers the idea of some 'semantic information service' and transfers the accessibility issue to the owner of such service. On the other hand, additional semantic information may be stored locally and thus foster distributed (information) networks. The question on how to manage such data (especially in terms of accessibility) remains a topic of research. Secondly, in order to store semantic information, an adequate syntax has to be selected. It is difficult to mention a universal solution for this purpose as any potential scheme has to be expressive

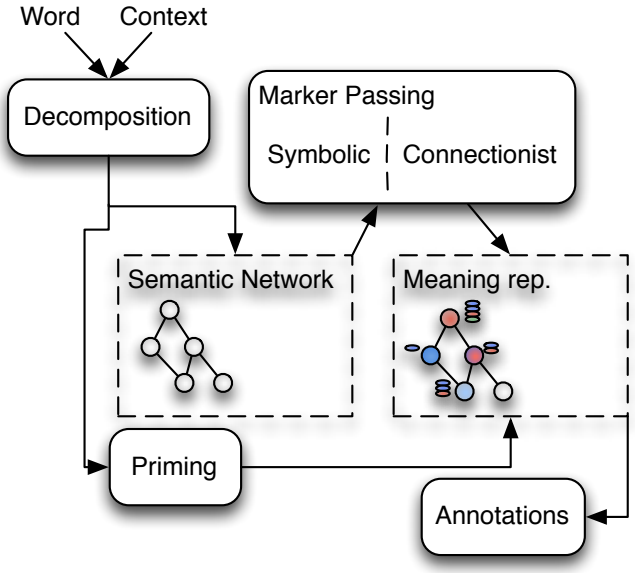


Fig. 1: Abstract approach to represent artificial meaning

on the one hand and domain specific on the other. Currently, there is much discussion on potential representation languages. Thirdly, the development of tools —especially of those tools which provide a graphical illustration of additional semantic data—is a research topic for itself. The problem of how to visualise semantic information becomes even more difficult with an increasing complexity of additional data. Finally, most of the examined approaches were not able to account for automated procedures. In addition to the question on how to realise automated procedures, the question on how much automatism is actually preferred, is widely discussed. Yet, having in mind, that manual annotation tools currently feature high level of sophistication, tools with automated support are likely to be the immediately next stage of evolution.

Thus the challenge of creating an automated annotation tool persists to date. Furthermore, to enable automatic processing the annotations should be computer readable (with a formal representation) so that future tools might use those annotations as information source. Even if a automatic annotation is reached, the possibility to manually influence the annotation should be given. This gives humans the possibility to correct wrongly created decompositions if, for example, a word sense disambiguation went wrong during the decomposition.

We identified the following components necessary to create an artificial representation of meaning that can be used to semantically annotate data.

As illustrated in Fig. 1 the self-explanation starts with building a model for the meaning of a word depending on the context by decomposing it. This leads to a semantic network representation (Ontology) of its denotation that represents the connectionist knowledge representation of meaning. Such a decomposition is done until semantic primes are reached, which need no further decomposition [65]. One challenge here is to select the right definitions of the word[14] from the utilised datasources to be used in the decomposition.

This semantic network is used to spread activation or pass markers through the network.[15] This is denoted by the different colours and markers in Fig. 1. The marker (represented as chips next to each node in the depiction) might carry symbolic information that steers the activation spreading. To be able to react to different markers, each node in the semantic network has a node interpretation function reflecting its behaviour. The node interpretation function inflects how the node processes incoming markers, how he passes outgoing marker on to other nodes and if he is activated. In this way, e.g., a “NOT” node passes its markers to the next node so that this one activates its opposites (in linguistic named antonyms). Since semantic relations like synonym and antonym relations have different meanings as well the relation interpretation function allows to specify how a relation passes on markers. In this way symbolic information like temporal logic can be encoded in the network. One challenge at this step is the amalgamation of the connectionist representation in the semantic network and the symbolic representation provided by a node and edge interpretation function.

During the activation through priming we can influence how the amalgamation of symbolic and connectionist representation of meaning is contextualised. By activating the right concepts out of the context, the marker passing will activate different nodes in the semantic network and thus contextualise the representation of meaning. Here the selection of parameters and concepts to activate is challenging. Finally we need an interpretation of the output of the marker passing to extract the meaning represented.

The automatic annotation then can be done by activating the word we want to annotate in the semantic network using the generated ontology of the marker passing for the annotation. If we want to annotate the word ‘Bank’ in a text discussing the financial crisis, the activation will have a stronger activation on ‘Bank’ as an financial institute then on the seating accommodation. This is because the priming will probably use words like money, accounting, currency or equivalents from the text during the activation. Thus the approach is able to annotate the text with context dependent meaning.

Regarding the proposed concept on how an automatic annotation component could be build, we want to extend the definition of Fähndrich et al. [66] of self-explaining system as follows:

Definition 1: A self-explaining system is able to create an internal knowledge representation of an unknown concept in a pragmatic manner through the use of external information sources and communicate the so-created meaning to other systems.

[14]This challenge is related to the word sense disambiguations and is one reason for the need of contextual information during the decomposition.

[15]Marker passing subsumes activation spreading since the classical activation spreading can be modelled with a marker that carries the activation level as numeric value.

Definition 1 has two parts: The first part requires the system to be able to explain new concepts to itself which means to create a denotation and a connotation in an manner that the system can reason upon this internal knowledge representation. The second part describes the ability to communicate this meaning to a other system in a manner that the other system is able to create its internal representation.

V. Conclusion

This work provides an overview on approaches, methods, and tools that support developers in comfortably viewing, editing and/or adding semantic information to relevant data. In doing so, we put particular emphasise on the inherent requirements of self-explaining systems. One important requirement here is the level of automation. Besides this and in order to classify the examined solutions several other properties were introduced. However, we focused our survey on approaches that automatically collect and add semantic information mainly at the applications runtime and distinguished the level of automation into four different increasing categories. To sum up, we can say that their are only a few solutions available that offer (semi)automatism capabilities. These solutions use, for example, learning algorithms to support users during the annotation process. Nevertheless, most of the examined solutions did not focus automation and we are far from fully automated annotations. In order to clarify the research progress here, we discussed the results of the survey. Substantiated by this discussion we revealed the limitations and formulated research challenges/questions that must be answered by the community. Here, beside the main question of how automatism can be realised, it might be interesting to discuss how much automatism is wanted respectively needed to create self-explaining systems and system components.

The results of the survey neglected the authors thought of an existing fully automated approach. With the goal to improve the state-of-the-art, we presented unsolved research challenges and plan to exercise some of them. Here, we will select and extend a fitting solution and try to increase the degree of automatism. However, at the very first, we want to discuss and formulate a reasonable and formalised definition for self-explaining systems.

VI. Acknowledgement

This work is being supported by the German government in the Bundesministerium für Wirtschaft und Energie (BMWi) Project: "Erweiterte und adaptive Elektromobilitätsdienste: Technologie, Entwicklung, Bereitstellung", FKZ: 16SBB007A.

References

[1] N. R. Jennings, "Building complex software systems why agent-oriented approaches are well suited for," *Communications of the ACM, Forthcoming*, vol. 44, no. 4, pp. 35–41, 2001.

[2] W. A. Gale, K. W. Church, and D. Yarowsky, "A method for disambiguating word senses in a large corpus," *Computers and the Humanities*, vol. 26, no. 5-6, pp. 415–439, Dec. 1992. doi: 10.1007/BF00136984. [Online]. Available: http://link.springer.com/10.1007/BF00136984

[3] M. Brodie, J. Mylopoulos, and J. W. Schmidt, *On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases, and Programming Languages*. Springer-Verlag GmbH, 1984.

[4] T. Berners-Lee, J. Hollenbach, K. Lu, J. Prebrey, E. P. d'ommeaux, and M. Schraefel, "Tabulator redux: Writing into the semantic web," Electronics and Computer Science, University of Southampton, Tech. Rep., November 2007.

[5] F. Lopez, "Overview of methodologies for building ontologies," in *KRR5*, V. Benjamins, B. Chandrasekaran, A. Gomez-Perez, N. Guarino, and M. Uschold, Eds., vol. 1999, August 1999, pp. 1–13.

[6] S. Braun and V. Zacharias, "Ontology maturing with lightweight collaborative ontology editing tools," in *ProKW 07*, N. Gronau, Ed. Potsdam, Germany: GITO, March 2007, pp. 217–226.

[7] O. Consortium, "Ontoweb: Ontology-based information exchange for knowledge management and electronic commerce," Vrije Universiteit Amsterdam (VU)-Coordinator Faculty of Sciences, Tech. Rep., 2008, last visited: 2012-06-06. [Online]. Available: v

[8] J. Cardoso, "The semantic web vision: Where are we?" *IEEE Intelligent Systems*, vol. 22, no. 5, pp. 84–88, September 2007.

[9] K. Suresh, J. Kumar Malik, N. Prakash, and S. Rizvi, "A case study on role of ontology editors," in *National Conference on Advancements in Information & Communication Technology (NCAICT)*, Allahabad, India, 2008.

[10] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 1, pp. 14–28, Dec. 2005. doi: 10.1016/j.websem.2005.10.002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1570826805000338

[11] N. Islam and Z. a. Shaikh, "Semantic web: Choosing the right methodologies, tools and standards," *2010 International Conference on Information and Emerging Technologies*, pp. 1–5, June 2010.

[12] Y. Ding and S. Foo, "Ontology research and development part 1 – a review of ontology generation keywords," *Journal of Information Science*, vol. 28, no. 2, pp. 123–136, 2002.

[13] A. Gomez-Perez and D. Manzano-macho, "Deliverable 1.5 : A survey of ontology learning methods and techniques ontoweb consortium," Madrid, Tech. Rep., 2003.

[14] L. Drumond and R. Girardi, "A survey of ontology learning procedures," in *In Proceedings of the 3rd Workshop on Ontologies and their Applications*, vol. 427, Salvador, Bahia, Brazil, October 2008, pp. 1–12.

[15] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, and S. W. Tu, "The evolution of protégé : An environment for knowledge-based systems development," vol. 58, no. 1, pp. 98–123, 2003.

[16] D. Huynh, S. Mazzocchi, and D. Karger, "Piggy bank: Experience the semantic web inside your web browser," in *ISWC*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, vol. 3729, pp. 413–430. ISBN 978-3-540-29754-3

[17] S. Handschuh, S. Staab, R. Volz, and L. Meyer, "Deep annotation for information integration," in *IIWeb-03*, S. Kambhampati and C. A. Knoblock, Eds., Acapulco, Mexico, August 2003, pp. 105–110.

[18] C. Bizer and T. Gauss. (2007) Disco - hyperdata browse. FU-Berlin. Last visited: 2013-07-20. [Online]. Available: http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[19] M. C. Schraefel, D. A. Smith, A. Owens, A. Russell, C. Harris, and M. Wilson, "The evolving mspace platform: leveraging the semantic web on the trail of the memex," in *ACM conference on Hypertext and hypermedia*, ser. HYPERTEXT '05. New York, NY, USA: ACM, 2005. ISBN 1-59593-168-6 pp. 174–183.

[20] J. Pozueco, D. Berrueta, L. Polo, J. E. Labra, and S. Fernandez. (2011) Rdfa developer. Javier Pozueco. Last visited: 2013-07-20. [Online]. Available: https://bitbucket.org/fundacionctic/rdfadev/wiki/Home

[21] K. Yee, K. Swearingen, and K. Li, "Faceted metadata for image search and browsing," in *SIGCHI conference on Human factors in computing systems*, 2003. ISBN 1581136307 pp. 401–408.

[22] O. Lassila, "Browsing the semantic web," in *International Workshop on Database and Expert Systems Applications, 2006. DEXA '06. 17th*, 2006, pp. 365 – 369.

[23] M. Butler, D. Huynh, B. Hyde, T. Berners-Lee, and M. R. (2006) Longwell project page. Simile. Last visited: 2013-07-20. [Online]. Available: http://simile.mit.edu/wiki2/Longwell

[24] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker, "Sig.ma: Live views on the web of data," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, no. 4, pp. 355–364, 2010.

[25] J. Kaizer and A. Hodge, "Aquabrowser library: Search, discover, refine," *Library Hi Tech News*, vol. 22, no. 10, pp. 9–12, 2005.

[26] D. Huynh, "Parallax and companion: Set-based browsing for the data web," in *ACM WWW Conference*, 2009. ISBN 9781595936547

[27] M. T. Pazienza, N. Scarpato, A. Stellato, and A. Turbati, "Din din ! the (semantic) turkey is served! from semantic bookmarking to knowledge management and," in *SWAP2008*, 2008, pp. 15–17.

[28] S. Krivov, R. Williams, and F. Villa, "Growl : A tool for visualization and editing of owl ontologies," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 2, pp. 54–57, 2007.

[29] J. Francis, M. Davies, and D. Mladenic, *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*, 1st ed. Berlin: Springer Berlin Heidelberg, January 2009.

[30] A. Kalyanpur, B. Parsia, E. Sirin, and B. Grau, "Swoop : A web ontology editing browser," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 2, pp. 1–20, June 2006.

[31] S. M. Paley, J. D. Lowrance, and P. D. Karp, "A generic knowledge-base browser and editor," in *AAAI97/IAAI97*, 1997. ISBN 4158593735

[32] F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks, "User-system cooperation in document annotation based on information extraction," in *EKAW02*, vol. 2473, no. October. Springer Verlag, 2002, pp. 122–137.

[33] D. McGuinness, R. Fikes, and J. Rice, "An environment for merging and testing large ontologies," in *KR*, 2000, pp. 12–15.

[34] M. Dzbor, E. Motta, and J. Domingue, "Opening up magpie via semantic services," in *ISWC*, ser. Lecture Notes in Computer Science, S. A. McIlraith, D. Plexousakis, and F. v. Harmelen, Eds., vol. 3298. Hiroshima, Japan: Springer-Verlag Berlin Heidelberg, 2004, pp. 635–649.

[35] A. Zhdanova and R. Krummenacher, "Community-driven ontology management: Deri case study," in *International Conference on Web Intelligence*, no. September, 2005, pp. 19–22.

[36] M. Fiorelli, M. T. Pazienza, S. Petruzza, A. Stellato, and A. Turbati, "Computer-aided ontology development: an integrated environment," in *NLPFrameworks 2010*. ELRA, 2010, pp. 28–35.

[37] D. Quan, D. Huynh, and D. R. Karger, "Haystack : A platform for authoring end user semantic web applications," in *ISWC*, 2003, pp. 738–753.

[38] IBM, "Emf ontology definition metamodel," last visited: 2012-06-06. [Online]. Available: http://www.eclipse.org/modeling/mdt/eodm/docs/articles/EODM_Documentation/

[39] L. Rutledge, M. Alberink, R. Brussee, S. Pokraev, W. Dieten, and M.Veenstra, "Finding the story: Broader applicability of semantics and discourse for hypermedia generation," in *ACM conference on Hypertext and Hypermedia*, 2003, pp. 67–76.

[40] D. Cameron, P. N. Mendes, A. P. Sheth, and V. Chan, "Semantics-empowered text exploration for knowledge discovery," in *IACM SE*. Oxford, MS, USA: ACM Press, April 2010. ISBN 9781450300643 pp. 1–6.

[41] A. Marchesini, "Morla project page," last visited: 2013-07-20. [Online]. Available: http://www.morlardf.net/index.php

[42] T. Eskridge, P. Hayes, and R. Hoffman, "Formalizing the informal: a confluence of concept mapping and the semantic web," in *Proc. of the Second Int. Conference on Concept Mapping*, 2006.

[43] B. Motik, "Reasoning in description logics using resolution and deductive databases," Ph.D. dissertation, Karlsruher Institut of Technology, 2006.

[44] I. Revelytix. Knoodle project page. Last visited: 2013-07-20. [Online]. Available: http://knoodl.com/

[45] L. Ceccaroni and E. Kendall, "A semantically-rich, graphical environment for collaborative ontology development in agentcities," in *iD3, Barcelona, Spain*, 2003.

[46] A. Farquhar, R. Fikes, and J. Rice, "The ontolingua server: a tool for collaborative ontology construction," *International Journal of Human-Computer Studies*, vol. 46, no. 6, pp. 707–727, June 1997.

[47] M. Rospocher, C. Ghidini, V. Pammer, SeraïňĄni, and Lindstaedt, "Moki: the modelling wiki," in *SemWiki 2009*, 2009, pp. 113–127.

[48] B. Fortuna and M. Grobelnik, "Ontogen: Semi-automatic ontology editor," in *Human Interface, Part II, HCII*, 2007, pp. 309–318.

[49] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, "MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup," in *Service-Oriented Computing – ICSOC 2013 Workshops*. Berlin, Heidelberg: Springer Berlin Heidelberg, Sep. 2002, pp. 379–391. ISBN 978-3-540-44268-4. [Online]. Available: http://link.springer.com/10.1007/3-540-45810-7_34

[50] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets, "Tabulator: Exploring and analyzing linked data on the semantic web," in *In Proceedings of the 3rd International Semantic Web User Interaction Workshop*, November 2006, pp. 1–16.

[51] U. Bojars, J. G. Breslin, V. Peristeras, and G. Tummarello, "Interlinking the social web with semantics," no. June, 2008.

[52] M. Hildebrand, J. van Ossenbruggen, and L. Hardman, "/facet: A browser for heterogeneous semantic web repositories," in *ISWC*, ser. Lecture Notes in Computer Science, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, Eds. Springer Berlin / Heidelberg, 2006, vol. 4273, pp. 272–285. ISBN 978-3-540-49029-6

[53] E. Oren, R. Delbru, and S. Decker, "Extending faceted navigation for rdf data," in *ISWC*, ser. Lecture Notes in Computer Science, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, Eds. Springer Berlin / Heidelberg, 2006, vol. 4273, pp. 559–572. ISBN 978-3-540-49029-6

[54] M. Stefaner, S. Ferré, S. Perugini, J. Koren, and Y. Zhang, *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*, ser. The Information Retrieval Series, G. M. Sacco and Y. Tzitzikas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 25. ISBN 978-3-642-02358-3

[55] F. Ciravegna, "$(lp)^2$, an adaptive algorithm for information extraction from web-related texts types of induced rules," in *In Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, Seattle, USA, August 2001.

[56] F. Ciravegna and Y. Wilks, "Designing adaptive information extraction for the semantic web in amilcare," in *Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications.* IOS Press, 2003, pp. 112–127.

[57] D. Griesi, M. Pazienza, and A. Stellato, "Semantic turkey: a semantic bookmarking tool (system description)," in *European conference on The Semantic Web: Research and Applications*. Innsbruck, Austria: Springer Verlag, 2007, pp. 779–788.

[58] M. T. Pazienza, N. Scarpato, and A. Stellato, "Stia: Experience of semantic annotation in jurisprudence domain," in *Legal Knowledge and Information Systems Jurix*, G. Governatori, Ed. IOS Press, 2009, pp. 156–161.

[59] R. M. MacGregor, "Using a description classi er to enhance deductive inference 1 introduction," in *IEEE Conference on AI Applications*, 1991, pp. 141—-147.

[60] P. Karp and V. Chaudhri, "A collaborative environment for authoring large knowledge bases," *Journal of Intelligent Information*, vol. 194, pp. 155–194, 1999.

[61] T. Nelson, "Complex information processing: a file structure for the complex, the changing and the indeterminate," 1965, pp. 84–100.

[62] L. Fischer, *The Perfect Swarm: The Science of Complexity in Everyday Life*. ReadHowYouWant, April 2010.

[63] J. Fähndrich, S. Ahrndt, and S. Albayrak, "Self-explaining agents," *Jurnal Teknologi (Science & Engineering)*, vol. 3, no. 63, pp. 53–64, 2013. doi: 10.11113/jt.v63.1955. [Online]. Available: http://www.jurnalteknologi.utm.my/index.php/jurnalteknologi/article/view/1955/1481

[64] S. Löbner, "Semantik. Eine Einführung," 2003.

[65] J. Fähndrich, S. Ahrndt, and S. Albayrak, "Formal Language Decomposition into Semantic Primes," *ADCAIJ: ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL*, vol. 3, no. 8, p. 56, Oct. 2014. doi: 10.14201/ADCAIJ2014385673. [Online]. Available: http://revistas.usal.es/index.php/2255-2863/article/view/ADCAIJ2014385673

[66] ——, "Towards Self-Explaining Agents," *PAAMS: Advances in Intelligent Systems and Computing*, pp. 147–154, 2013.

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 25–30

DOI: 10.15439/2015F380
ACSIS, Vol. 6

The Serialization of Heterogeneous Documents

Peter John Hampton, William Blackburn, Hui Wang
Artificial Intelligence and Applications Research Group
Ulster University, Jordanstown
United Kingdom, BT37 0QB
Email: hampton-p1@email.ulster.ac.uk
{wt.blackburn, h.wang}@ulster.ac.uk

***Abstract*—Tasks involving the analysis of natural language are typically conducted on a corpus or corpora of plain text. However, it is rare that a document is unstructured and freeform in its entirety. Documents such as corporate disclosures, medical journals and other knowledge rich archive contain structured and loosely-structured information that can be used in a variety of important text mining tasks. In this paper we propose a syntactical preprocessing architecture to serialize presentation-oriented documents to a machine readable format that aspires to preserve the document structure, contents and metadata. We introduce a hybrid pipeline architecture, discussing the various processes and the future research direction that could potentially lead to a holistic representation of heterogeneous documents.**

I. Introduction

KNOWLEDGE mining researchers and practitioners have been implementing techniques to aid and enact decision-making from knowledge discovery tasks. However, various challenges restrict the computational understanding of language found in such documents as analysis has traditionally focused on plain text (1; 2; 3). This paper proposes a hybrid preprocessing architecture for preserving a documents contents in its entirety and converting selected entity classes to their canonical form, enabling deeper analysis.

Although there are arguably many documents of interest, we focus the attention of this paper on corporate disclosures, specifically interim financial reports (10-Qs) due to the depth of knowledge and the complexity of their composition. We demonstrate, at a high level, a multistage architecture in Fig. 1 that combines both statistical and rule based approaches for serialization to preserve the documents structure and content.

Diverse developments in Information Retrieval methodology have been made over the past two decades, which could make it sufficiently easier to represent unpredictable document formats and associated contents. We describe related work and motivations behind this research in Section II. Section III analyzes the document structures of five company interim disclosures. In Section IV we compare various mainstream data serialization formats while concluding the advantages and disadvantages among them. Section V describes the pipeline components depicted in Fig. 1 in substantive detail, breaking down each process into a set of processes. The paper concludes in Section VI which sets a future direction for our research.

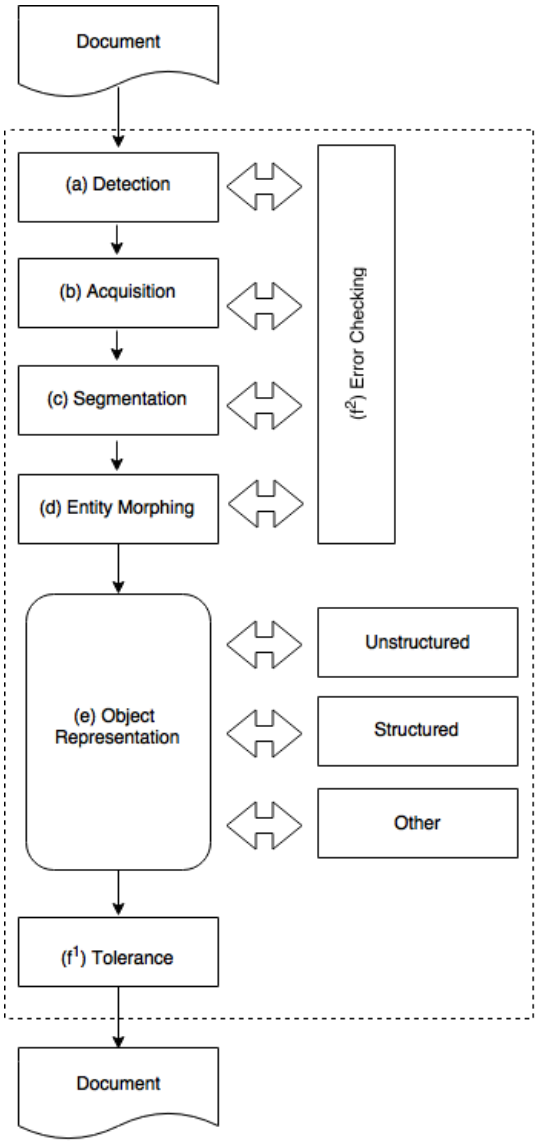


Fig. 1. The Hybrid API Architecture described in this paper for the syntactic representation of heterogeneous documents. At a high level, the depicted analysis pipeline is broken down into 6 stages: Detection (a), Acquisition (b), Segmentation (c), Entity Morphing (d), A cache object for document representation (e) and a duel error handling module (f). A document is introduced into the pipeline and subject to numerous decomposition stages.

II. Related Work

Early attempts at analyzing structured data in heterogeneous documents include the Douglas *et al* study (4) which represented tabular data using spaces in plain text. These researchers presented an analysis of table layout and the associated linguistic characteristics. However, wider linguistic complexities are presented in plain text representation of tables. This can include complex header parsing, redundant or null cell representation and wider linguistic complexities. We aim to extend their problems of characteristic syntax and apply a set of transformations to the contents.

More recent work includes the Clark & Divvala research (5) which aims to discover a holistic view of the document, achieving a deeper *semantic* understanding of articles such as academic computer science papers by placing emphasis on figures such as charts. Likewise, the data found in corporate disclosures tend to be multifaceted in nature, that being a mix of unstructured, loosely-structured, unstructured text and miscellaneous data. Their open-source solution agnostically parses a document and locates the areas wherein figures or tables could reside by reasoning about the empty regions within that text, achieving success due to it's relaxed formatting assumption.

III. Background & Analysis

In this paper we refer to freeform text, that is sentences with no predictable syntactical structure as *unstructured* content. In turn, we refer to information in tabular form as *structured data*. Although the term 'loosely-structured data' is typically used to refer self-describing data (6; 7), we use it in this paper as a means of classifying information that is not presented in a structured or unstructured form. This can take the form of bullet points, footnotes, images etc.

We analyzed a random sample of five interim statements published by different software companies between July and December 2014 as listed on the British Alternative Investment Market (AIM). The results portrayed in Fig. 2 show that although the majority of content is unstructured in nature, the sample set had structured and loosely-structured data accounting for 22% to 65% in terms of token count.

Fig. 3 on the other hand shows that although unstructured, free form text and miscellaneous data increased with page count, structured and loosely-structured data could be considered random and unpredictable. The share of structured and loosely-structured information is, in our opinion, substantial and shouldn't be left out of text mining tasks due to the risk in knowledge loss. We propose syntactically serializing the document in a machine readable format whose schema is flexible enough to adapt to unpredictable content and volatile formatting. It is clear however from this preliminary analysis of the small data set that if it is possible to serialize documents, a format would require several characteristics to make this possible such as flexibility, system interoperability, etc. We review multiple data serialization formats in the next section and select the most appropriate for serializing multifaceted documents written by specialist humans.

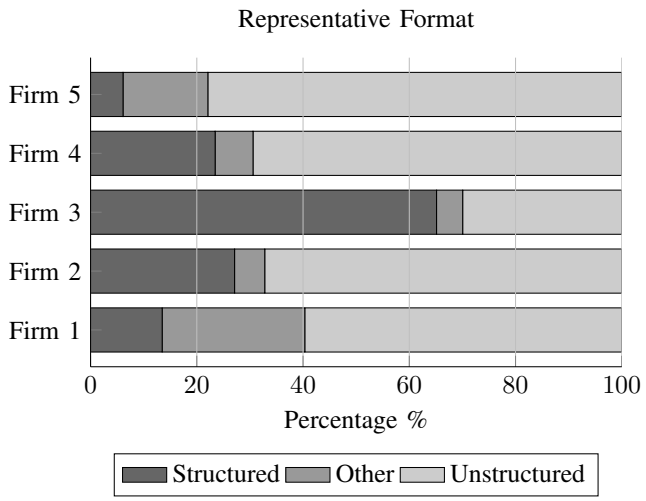


Fig. 2. Five Interim statements broken down to their representative cluster. This initial study found there was no predictable cluster share of the documents studied.

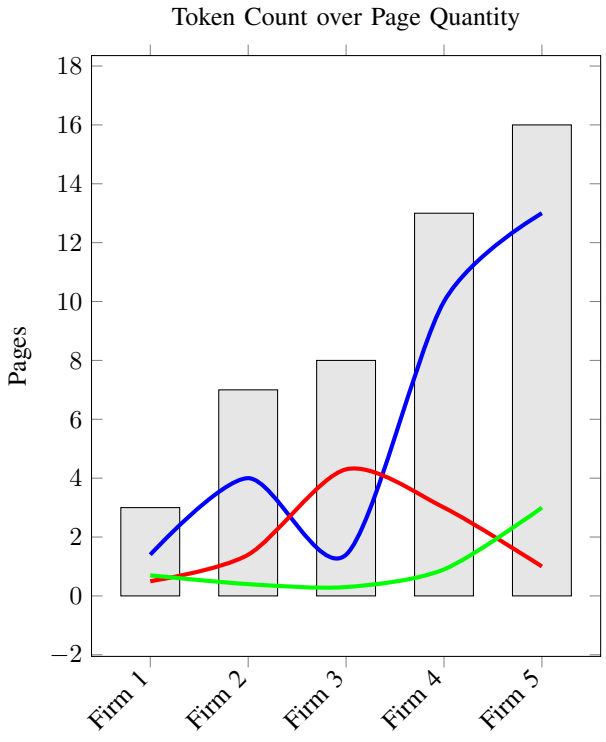


Fig. 3. The change in token count over page count. The results show that although unstructured content increases with page count, it can be deemed unpredictable the share of structured, loosely structured and other miscellaneous data will represent of the over all document. The blue line shows the change in unstructured data wheras The red line shows changes in structured data and The green line shows the changes in other data formats.

TABLE I
THE DATA TYPES OF JAVASCRIPT OBJECT NOTATION

Data Types	Example	Description
Numeric	1, -5, 0.9, 34543987584	Can represent a signed integer or floating-point number and may use exponential E notation.
String	“Jack”, “Jill”	A sequence of characters. This data type supports 8, 16 and 32 bit unicode transmission formats.
Boolean	true or false	Either True or False values.
Array	[0, “Jill”, [1, 2], null]	An ordered list of all the data types in this table including null.
Object	“names”{ “name”: “Jack”}	An unordered set of key-value pairs, where the keys must be specified explicitly in a String format.
None Type	null	A vacant value often used as a placeholder.

IV. SERIALIZATION FORMATS

Data Serialization is the process of translating data structures into a sortable format to be loosely reconstructed at a later time if required. We focus our attention on formats that are data oriented rather than document markup oriented. In the upcoming subsections we review XML, JSON and YAML.

A. XML

XML (eXtensible Markup Language) is probably one of the most popular document serialization formats since the rise of Web 2.0, a milestone in web evolution that democratized content, converting a generation of information consumers into content creators. XML is described by the World Wide Web Consortium as a flexible user specified markup scheme, whose elements are not subject to formatting rules (7). One advantage XML has is that it is human readable and can be converted to various serialization formats using XLST (eXtensible Stylesheet Language Transformation) or an independent parser.

B. JSON

JSON (JavaScript Object Notation) is a relatively newer serialization format compared to XML, which is lightweight, thus lower processing overhead (8). It is regarded as a lightweight alternative to XML, which can preserve the native data type of the selected entity and can be used for server parsing (9; 10). We describe an overview of JSON data types in Table I.

C. YAML

YAML (YAML Ain’t Markup Language) is a superset of JSON that offers an alternative, user friendly syntax by replacing various nested delimiters, such as list braces, object colloquially and quote marks with whitespace. YAML supports a powerful feature that we believe could benefit many natural language processing researchers called anchoring, which enables embedded object referencing and can handle relational information similar to a traditional SQL database (11).

D. Discussion

Due to its popularity, XML seems an attractive choice. However to query an XML, XPATH (XML Path Language) is required to parse the document and would be considered inefficient compared to newer standards. Further we believe it could prove difficult to train a machine to autonomously parse the semi-structured nature of XML document. Adopting JSON on the other hand, a specified path could be explicitly declared or discovered with ease due to the tree like structure. After reviewing numerous studies and experimentation, we chose to rule out YAML in it’s current state[1] as it typically requires an external parser and we found that when automatically generating YAML documents there was possible corruption. However, future experimentation with YAML could yield an interesting study as it is much more verbose and is being used in various innovative ways.

JSON compared to XML has a steeper learning curve and debatably a much more complex syntax. The comprehensive Eriksson-Halberg study surveyed performance and the syntax of JSON and YAML and found JSON implementations to be ‘*many times faster than YAML for both serialization (dumping) and deserialization (loading)*.(12)’ The researchers concluded that the complexity behind processing YAML formats to be the reason for this. We feel that due to the level of expressivity and efficient performance of JSON that we should adopt it, for the current time being, to represent document inputted to our pipeline.

V. PIPELINE ARCHITECTURE

In Fig. 1, we showcased and briefly discussed our high-level pipeline architecture for serializing documents with an unpredictable structure and format at a high level. Here, we describe the various stages an inputted document goes through to be serialized. As corporate disclosures, medical journals and academic papers are commonly published in Portable Document Format (PDF), we describe a PDF as the input to this propositional system.

A. Detection

The document format is first determined by the MIME type (Multipurpose Internet Mail Type). This phase inserts the content type header along with the document to the Acquisition phase (b) for the documents conversion and decomposition.

B. Acquisition

Disclosures released by a company are often published in Portable Document Format (PDF). This *presentation over*

[1]At the time of publication, the latest version of YAML was v1.2

content oriented format presents a number of problems for natural language acquisition and analysis. Further, the document's structure is unpredictable and can change between structured, loosely-structured or free form texts at any time. Therefore, we first propose an introduction stage dedicated to the decomposition of the document, which we call the *Acquisition and Segmentation* phase of our pipeline shown in Fig 4.

Once the pipeline has accepted the document, two concurrent processes must take place. The first process aims to extract the document metadata. This metadata can include information such as creation date, modified date, author name(s), publisher geo-coordinates and other miscellaneous information (13).

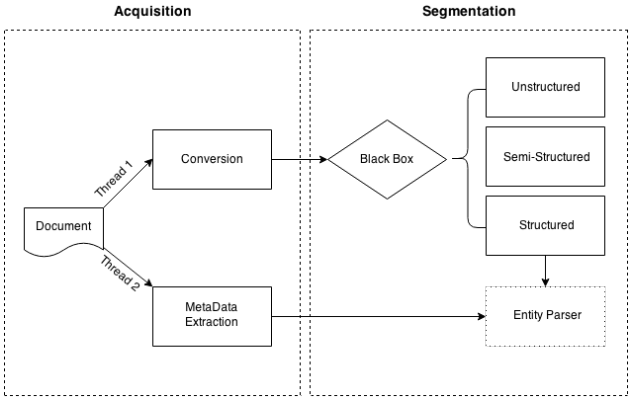


Fig. 4. The Acquisition Process: The detected document format is accepted bu the pipeline. One process extracts the metadata (if any) and the second process clusters the document into one of three categories.

C. Segmentation

The segmentation process converts the document into an HTML (HyperText Markup Language) format, assuming the content type header is not already in HTML format, preserving bullet points and table formats, the latter for example by placing null values in redundant cells. The source document is first converted to HTML in order to traverse semi-structured nodes in the document. To categorize the unstructured, loosely-structured, structured and other miscellaneous information, we propose a black box clustering mechanism to classify innumerable sections of the document that the set of token sequences belong to. During this phase, we also assign various reference data to the section, which includes the page number, the section number, the paragraph index, and index of the sentence in that paragraph, to somewhat preserve the presentational elements of the document.

D. Entity Morphing

Due to the unpredictable and creative use of language within corporate disclosures, the need to manipulate tokens became apparent at early stages of this research. This component described is a series of objects that manipulate text into a common reusable and storable format in the form of objects. We narrow the research direction to seven named entity types, which can be efficiently remodeled using pattern matching techniques such as regular expressions. These entity types are *monetary values, percent, decimals, durations, times, dates and miscellaneous quantity values* with examples provided in Fig 5.

$(\$1.1b \rightarrow \{'\$', 1100000000\})$
$(\$4.7million \rightarrow \{`\$', 4700000\})$
$(6.3p \rightarrow \{'GBP', 0.063\})$
$(12thMarch2015 \rightarrow \{'20150312000000'\})$
$(€66,000,000 \rightarrow \{'€', 66000000\})$
$(5.2\% \rightarrow \{0.052\})$
$(-2perc \rightarrow \{-0.02\})$

Fig. 5. Parsed Tokens: examples depicting an input x and output y where (.) represents an input to the annotator object's helper function. $(x) \rightarrow y$

We advocate the removal of PERCENT objects by converting all percentages found in the text to a decimal format. It was found that performing calculations on the extracted entities was possible if presented in a decimal format, a technique we plan to explore further in future research.

E. Object Representation

We refer to unstructured content as content that has no predefined data-model and has an unpredictable structure, often mass text that can contain various objects such as money, time, quantities, and so on. We serialize the sentences into an array of JSON objects, maintaining the various indexes discussed in Section A of the pipeline. We use the Punkt Tokenizer in the NLTK as described by Bird (13) which is a model trained using a unsupervised machine learning algorithm for sentence boundary detection. However, various popular tokenizers could also prove appropriate for this task such as the Stanford Tokenizer or TrTok which prove effective when parsing messy web text data. We provide a serialization example of an unstructured paragraph in 5.1.

Example 5.1 (Unstructured Paragraph Example):

```
{
    "sentences": [{
        "paragraph_index": 1,
        "sentence_index": 1,
        "cluster_index": 1,
        "page_index": 1,
        "sentence": "This is a sentence."
        },
        {
        "paragraph_index": 1,
        "sentence_index": 2,
        "cluster_index": 1,
        "page_index": 1,
        "sentence": "...also a sentence."
    }]
}
```

Khusro *et a*l (15) note that the *detection, extraction and annotation* of tables within heterogeneous documents have been quite a significant research problem in Information Retrieval for many years. The structured serialization stage of the hybrid pipeline aims to serialize complex and unpredictable tables into JSON format. Tables are information rich data stores, which contain a lot, often audited factual or objective information. We show how a parsed table (Table II) would be serialized in JSON in Example 5.2. This is possible using a very carefully programmed set of rules. This white box approach has proved effective for our small sample of documents but may need to be extended with intelligent based processing as our architecture scales and formats become increasingly complex.

TABLE II
FDP.L PROFIT & LOSS EXCERPT DATED NOVEMBER 2014

	6 Months Ended, 31 August 2014	**6 Months Ended, 31 August 2013**
	£'000	**£'000**
Revenue	37,507	34,381
Cost of Sales	(27, 606)	(25,313)
Gross Profit	9,900	9,068

Example 5.2 (FDP.L Table Serialized):

```
{
    "profit_loss": {
        "revenue": {
            "2014-08": 37507000,
            "2013-08": 34381000
        },
        "cost_of_sales": {
            "2014-08": -27606000,
            "2013-08": -25313000
        },
        "gross_project": {
            "2014-08": 9900000,
            "2013-08": 9068000
        }
    }
}
```

F. Error Checking & Tolerance

We implement two logic based error checking modules depicted in Fig. 1 as f_1 and f_2 respectively. The error handling processes verifies the accuracy and corrects mistakes made by the various concurrent processes. Scenarios covered by the error checker include:

- The document format has been correctly identified.
- Verify an objects reference data.
- Cast string named entities to their native format.
- Testing the structure of the tree-like output.

VI. FUTURE WORK

In this initial study we have proposed a novel architecture to serialize and represent a document and its contents for further analysis. We feel there is still scope for vast improvement and research into converting presentation oriented documents into a machine readable format that a human can easily debug. First, with the rise of the semantic web and linked data, we believe that extending our serialization model to JSON-LD (JavaScript Object Notation for Linked Data) may be appropriate. Secondly, there is potential to cluster knowledge together as in *Example 6.1* to solve various interoperability and distribution problems within Big Data systems.

Example 6.1 (Multiple Sources Serialized):

```
{
    "id": 1,
    "name": "Company X",
    "collection": {
        "disclosures": { ... },
        "social_media": { ... },
        "company_news": { ... },
        "media_news": { ... },
        "stock_price": { ... }
    }
}
```

Finally, many messages published over the web are transmitted natively in a JSON format and could prove appropriate to build up a profile of different sources into a single object for efficiency. We believe from this early study that this would make Big Data analysis on these documents much more efficient and manageable (15).

REFERENCES

[1] Comeau, D. C., Liu, H., Doğan, R. I., & Wilbur, W. J. (2014). Natural language processing pipelines to annotate BioC collections with an application to the NCBI disease corpus.

[2] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).

[3] Liu, M., Xu, W., Ran, Q., & Li, Y. (2015). Using Natural Language Processing Technology to Analyze Teachers' Written Feedback on Chinese Students' English Essays.

[4] Douglas, S., Hurst, M., & Quinn, D. (1995). Using natural language processing for identifying and interpreting tables in plain text. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (pp. 535-546).

[5] Clark, C., & Divvala, S. Looking Beyond Text: Extracting Figures, Tables and Captions from Computer Science Papers.

[6] Ding, L., Zhou, L., Finin, T., & Joshi, A. (2005). How the semantic web is being used: An analysis of foaf documents. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on(pp. 113c-113c). IEEE.

[7] Li, X., Li, F., & Chen, X. (2015, April). Distributed GIS framework design based on XML and Web Service. In 2015 International Conference on Intelligent Systems Research and Mechatronics Engineering. Atlantis Press.

[8] Hwang, C. G., Yoon, C. P., & Lee, D. (2015). Exchange of Data for Big Data in Hybrid Cloud Environment.

[9] Niu, Z., Yang, C., & Zhang, Y. (2014). A design of cross-terminal web system based on JSON and REST. In Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on (pp. 904-907). IEEE.

[10] Smith, B. (2015). Creating JSON. In Beginning JSON (pp. 49-67). Apress.

[11] Ben-Kiki, O., Evans, C., & Ingerson, B. (2005). YAML Ain't Markup Language (YAML^{TM}) Version 1.1. yaml. org, Tech. Rep.

[12] Eriksson, M., & Hallberg, V. (2011). Comparison between JSON and YAML for data serialization. The School of Computer Science and Engineering Royal Institute of Technology.

[13] Bird, S. (2006). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). Association for Computational Linguistics.

[14] Smutz, C., & Stavrou, A. (2012, December). Malicious PDF detection using metadata and structural features. In Proceedings of the 28th Annual Computer Security Applications Conference (pp. 239-248). ACM.

[15] Khusro, S., Latif, A., & Ullah, I. (2014). On methods and tools of table detection, extraction and annotation in PDF documents. Journal of Information Science.

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

DOI: 10.15439/2015F1
ACSIS, Vol. 6

Processing Imprecise Database Queries by Fuzzy Clustering Algorithms

Anna Kowalczyk-Niewiadomy
Technical University of Lodz
Lodz, Poland
anna.kowalczyk-niewiadomy@p.lodz.pl

Adam Pelikant
Technical University of Lodz
Lodz, Poland
adam.pelikant@p.lodz.pl

***Abstract*—Nowadays database management systems are one of the most critical resources in every company. Despite advanced possibilities of SQL, relational database management systems do not support flexible query conditions.**

Main assumptions of this work were two facts. First, that real data not representing random distribution (white noise), but have natural trend to granularity. The second one, that in everyday contacts we do not using strict defined conditions. The second feature lead us to use fuzzy logic which closer representing natural communication. First gives us opportunity to automatically construct functions defining membership to discrete groups based only on data distribution.

The problem of extending database systems with natural language expressions is a matter of many research centers. The basic idea of presented research is to extend an existing query language and make database systems able to satisfy user needs more closely. This paper deals mostly with gaining imprecise information from relational database systems. Presented concept is based on fuzzy sets and automatic clustering techniques that allow to build membership function and fuzzy queries. Thanks to applied solutions, the relational database system is more flexible, and similar to natural way of communication.

***Index Terms*—fuzzy logic, fuzzy sets, fuzzy set theory, fuzzy systems, fuzzy clustering, FSQL, fuzzy queries, imprecise queries.**

I. Introduction

THE SECOND half of the twentieth century ushered in rapid development of technology, especially in information technology. The growing demand for storing and processing huge data sets has resulted in evolution of database management systems. Those systems are designed to ensure the cohesion and safety of stored data and their principal objective is to search large data sets efficiently. Despite advanced possibilities of SQL, it is restricted to the precise communication only. In most business applications, querying precise values or using standard sharp relationships and traditional methods of data aggregation is absolutely sufficient. However, in some cases a standard SQL language, which is based on three-value logic, is not flexible enough. For example, if one is looking for cheap accommodation, or wish to buy a house that costs around €100 000, it is impossible to get results that will satisfy him, by means of traditional precise query language. Both of presented queries use natural language features that are used in everyday life. Traditional SQL is not feasible to build a query that supports such imprecise expressions. Imprecision in such context should not be seen as a drawback, but on the contrary, it allows expressing true needs, preferences and evaluation.

II. The Current State Of Knowledge And Relevance Of Research

Over the years, traditional methods of searching for information based on the precise conditions are more often replaced by methods based on fuzzy logic elements. The first fuzzy query language was presented by Takahashi in 1991 [17]. Two years later he published the full theory of two languages: calculus query language and fuzzy algebra query language [18]. In the eighties the problem of fuzzy database were investigated by: Zamenkowa [23], Chang Ke [7], Buckles and Petry [4], [5].

In the early nineties, thanks to the rapid information technology development, we could notice first implementations of fuzzy query systems. In France, P.Bosc and O. Pivert designed SQLf – fuzzy language which allows getting imprecise information from database [3].

Almost at the same time, in Poland professor S. Zadrożny and professor J. Kasprzyk from Systems Research Institute at Polish Academy of Science (PAS), presented FQUERY system for Access database. FQUERY consists of tools that enable user building queries with fuzzy values, relations, linguistic modifiers

and quantifiers. Currently, scientists who were mentioned above, work on linguistic summaries of databases problems and publish their achievements together with P.Bosc and O. Pivert from Malaga [4].

After the year 2000, there were presented newer solutions based on today's leading database management systems. For example, the research team led by Dr. Jose Gomes Galino of the University of Malaga, have developed system FSQL for Oracle 7/8, available on the Internet [12]. In Poland, Technical University of Poznan [9] and Silesian University of Technology [11] designed their own fuzzy systems SQLf and Fuzzy Logic Management System.

Despite many implemented systems for query languages, this branch of science is still being investigated and requires extensions to cope with the growing demands. So far, solutions based on the fuzzy sets theory contain strong constraints on the design stage of fuzzy sets. The developed systems are based on rigidly defined membership functions, and therefore require cooperation with expert's knowledge.

The basic idea of our research is to redesign a fuzzy structured query language system by extending traditional SQL (Oracle 11g) with condition definition similar to natural language. Due to the fact that few fuzzy query systems already exist in Polish as well as foreign research centers, it is worth to emphasize that the innovative element of this work is development of universal, multi-dimensional algorithms, which automatically generate fuzzy sets, based on the real data distribution. There is no need to use expert knowledge while original algorithms based on fuzzy clustering methods are implemented. In addition, conducting a comprehensive analysis of standardization issues and the labeling process enabled implementation of an intelligent module responsible for the allocation of labels according to the automatically generated fuzzy sets. There are some arguments to build such solution. The work of branch experts generates additional high costs. In many cases the meaning of label changes as a result of data distribution changes. For example prices of apartments usually grow up, so that meaning of cheap and expensive flat changes as well. Additionally in times of crisis prices can rapidly drop. Any of these states requires the help of experts, which can be avoided if the proposed solution is used. The approach (in more detail discussed in Chapter VI) is a completely new idea in the fuzzy SQL language issues.

III. Fuzzy Clustering Algorithms

Data clustering is a process of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Each data point belongs to a cluster of a degree of membership grade. This paper reviews three of the most representative clustering techniques: Fuzzy C Means, Fuzzy C Medoids clustering and Mountain clustering. All techniques were implemented and tested against automatic fuzzy sets generation problem. Fuzzy clustering methods together with the author's algorithm and the trapezoidal membership function allowed generating fuzzy sets based on the actual data distribution.

A. Equations

Let X be a set of n patterns described by $X = \{x_1, x_2, \dots, x_n\}$. Let c be an assumed number of clusters. $C = \{c_j | \ 1 \le j \le c\}$ is the set of centers. The notation $u_{ij}(1 \le i \le n, 1 \le j \le c)$ indicates the degree of membership of the i-th sample to the j-th prototype. The membership matrix *U* is limited to values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity (1).

$$\sum_{j=1}^{c} u_{ij} = 1; \ 1 \le i \le n \tag{1}$$

The Fuzzy C-means method was proposed in 1973 by Dunn [10] and modified in 1981 by Bezdek [1]. The algorithm is based on clusters search in a data set, such that an objective function (2) of distance measure is minimized. The squared distance is weighted by the m-th power of the membership in cluster j.

$$J_m(U,c) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^m \left\|x_i - c_j\right\|^2 \tag{2}$$
$$1 \le m \le \infty$$

Distance measure can be expressed by one of specific forms general Minkowski (3) norm [20]

$$d_n(x_i, x_j) = \left(\sum_{k=1}^{d} \left(\left|x_{i,k} - x_{j,k}\right|\right)^n\right)^{\frac{1}{n}} \tag{3}$$

One should remember that the result of clustering depends on kind of selected metric. The (4) and (5) are mandatory conditions for equation (2) to reach its minimum.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \tag{4}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{5}$$

The algorithm works iteratively through the preceding two conditions until there is no more improvement noticed. FCM calculates cluster centers and the membership matrix U using the steps presented at figure Fig 1.

The main advantage of the FCM algorithm is high performance and low hardware requirements. Unfortunately, this algorithm has three major drawbacks. First, the final distribution of objects between clusters strongly depends on the assumed number of clusters. Second, the performance of FCM depends on the initial membership matrix values. It is advised to run the algorithm for several times, every time starting with new values of membership grades for data points. Third, the algorithm is sensitive to disrupted data (e.g. singular point).

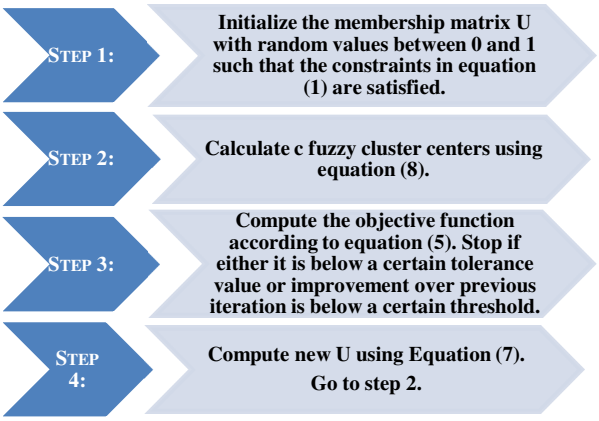


Fig 1. FCM – processing steps.

In order to solve this problem, instead of calculating mean we can search the most centrally located cluster point called medoid. In this way, the occurrence of the singular point in the cluster will not cause significant disruptions.

B. *Fuzzy C-Medoids Clustering (FCMdd)*

Fuzzy C-Medoids Clustering [8], relies on the basic idea of Fuzzy C-means clustering (FCM) with the difference of calculating cluster centers. The change has a significant influence on the efficiency of the algorithm. Instead of searching for means (calculated as a simple arithmetic formula) we need to process several steps over the neighbor points to find medoids. The improvement (minimization) of the criterion function (6) is much more complex and expensive. The notation $r(x_i, v_j)$ indicates dissimilarity between the x_i sample and v_j medoid.

$$J_m(V;X) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} r(x_i, v_j) \qquad (6)$$

Membership matrix (u) is calculated according to (7):

$$u_{ij} = \frac{\left(\frac{1}{r(x_j,v_i)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{r(x_j,v_k)}\right)^{\frac{1}{m-1}}} \qquad (7)$$

Figure 2 presents basic steps of the FCMdd algorithm.

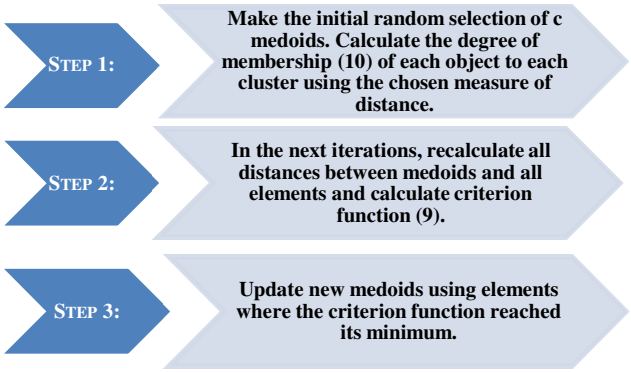


Fig 2. FCMdd-processing steps

C. *Mountain Clustering*

The mountain clustering method, proposed by Yager and Filev [22] is a simple and effective algorithm based on a density measure called the mountain function. It is based on three main steps. The first one involves forming a grid on the data space, where the intersections of the grid lines constitute the potential cluster centers. The second step entails construction of a mountain function representing data density measure. The height of the mountain function at a point $v \in V$ is equal to (8):

$$m(v) = \sum_{i=1}^{N} e^{-\left(\frac{\|v-x_i\|^2}{2\sigma^2}\right)} \qquad (8)$$

where x_i is the i-th data point and σ is an application specific constant. The third step involves selection of the cluster centers by sequentially reducing the mountain function. This is done by modification of the mountain function to the form represented by equation (9):

$$m_{new}(v) = m(v) - m(c_1)e^{-\left(\frac{\|v-c_1\|^2}{2\beta^2}\right)} \qquad (9)$$

IV. CLUSTER VALIDATION PROBLEM

The problem of data clustering is quite complex, what is mainly caused by wide potential of methods usage. Depending on the situation there is a need to use different types of algorithms, so it is difficult to impose a universal method. One of the main subjects in data clustering is evaluation of the result of clustering algorithms (cluster validation). More precisely, the

cluster validation problem is to find an objective criterion to determine how good a partition generated by a clustering algorithm is. Since most clustering algorithms require a pre-assumed number of clusters, a validation criterion to find an optimal number of clusters would be very beneficial.

The first validation was associated with the FCM partition coefficient proposed by Bezdek [1], defined by (10):

$$I_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{10}$$

To produce a better clustering performance we find optimal cluster numbers for $\max_{2\le c\le n-1} I_{PC}$.

Partition entropy was also proposed by Bezdek for the Fuzzy C-Means algorithm and it is defined by the following equation (11):

$$I_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_2 u_{ij} \tag{11}$$

To produce better clustering performance we find optimal cluster numbers for $\min_{2\le c\le n-1} I_{PE}$.

In 1991 Xie and Beni [19] proposed a validation index based on compactness and separation defined as (12):

$$I_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\|x_j - v_i\right\|^2}{n \min\limits_{i\ne j} i,j \left\|x_j - v_i\right\|^2} \tag{12}$$

In 2011 Rubio, Castillo and Melin [16] compared the most commonly used indices such as I_{PC}, I_{PE}, I_{XB} and proposed its own (13) proving its greatest effectiveness.

$$I_{RCM} = I_{MPE} + D_M \tag{13}$$

where

$$I_{MPE} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^2 \log_2 u_{ij} \tag{14}$$

$$D_{M_k} = \sum_{\substack{i,j=1 \\ i\ne j}}^{k} \left\|M_i - M_j\right\|^2, k = 1,\dots,c \tag{15}$$

During the research **all presented above validity indexes were implemented** and used in the process of fuzzy sets generation. In order to evaluate the quality of the indexes we used multi-dimensional data structures from the actual Machine Learning Repository maintained by the Center for Machine Learning and Intelligent Systems, University of California, Irvine [6]. Detailed experiments revealed that validity indexes based on compactness and separation are most effectiveness, that is why in our fuzzy SQL system the Rubio and Xie-Beni indexes were most important.

V. FUZZY SQL SYSTEM DESIGN

The main idea of our research was to design and implement system, which extends traditional SQL with condition build in the way similar to natural language. The great difference between proposed solution and already existing similar systems is fully automated generation of membership functions and fuzzy sets [13]. In addition to this, the automatic labeling module is also novelty[14], [15]. The project consists of three main modules described below.

The basic problem of testing is to gather representative data set. It should provide the actual distribution of the data and their continuous growth. The tests were carried out on several public available datasets. Finally we decided to build and positioned the dedicated WWW portal with tutorials for some programing and database platforms. Because the data set is one of the most important factor for all tests realized for implementing algorithms we decided to use selected Google Analytics statistics of the website traffic (fig. 3):

- number of visits (total and unique users);
- number of page views;
- percentage of rejections (input of 1 page views);
- average time spent on the site;
- loyalty (percent of new visits).

Fig 3. Exemplary overview of Google Analytics for defined data period.

Results describing portal users activity store in Google Analytics were exported to database server. This step concerns gathering an input data, and preparing database model for the main processing are presented at Fig 4.

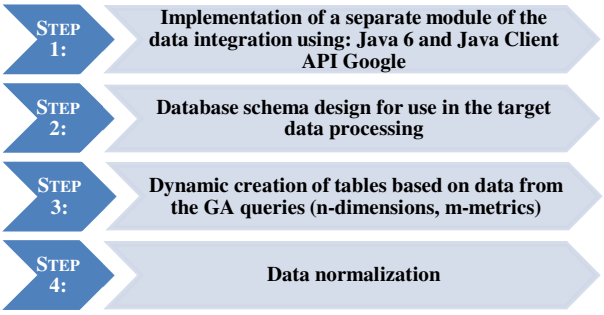


Fig 4. Data preparation diagram.

The second element of created system is a set of tools responsible for fuzzy queries processing consisting of the following components:

- query parser, which decomposes the fuzzy query into understandable by the system and database elements;
- fuzzy sets generator – a complex module using fuzzy clustering methods, novel fuzzy sets algorithms, validity indexes, T/S norms implementations etc.;
- labels assignment process – each fuzzy set is addressed by appropriate label or labels;
- fuzzy conditions, operators and aggregate functions executor.

The third element of the project – the client application is responsible for construction fuzzy queries (tree view and text) and presentation of the results.

A. *Data preparation*

In order to collect a sufficient data set, there was a need to prepare (design, implement and publish) an exemplary website which provides tutorials for web graphics (3dStudioMax, After Effects, Photoshop). Website traffic statistics are a good source of natural input data. In our project the Google Analytics (GA) was used as a statistical data warehouse. GA is a free, online tool mainly used to analyze websites statistics. GA is a powerful tool with variety of functionalities accessible via web browser and what was very useful via Java API.

One of the most important elements of the research was to prepare software in order to communicate easily with the GA. By use Java programming language, Java GA API (JGA), and JDBC the integration with the Oracle database was provided. Figure 4 presents basic GA integration steps.

It is worth to emphasize that data prepared for fuzzy clustering algorithms are normalized to the <0, 100> range in order to eliminate the problem of scale, negative values and to ensure the integrity of the generated fuzzy sets. Standardization refers to both input data as well as a range of labels. In this way, it is possible to assign labels by percentage match in order to avoid context dependency. For example query using "high temperature" expression in context of the weather, boiling water or melting metals is completely different. Data normalization eliminates this problem.

GA enables lots of important statistics about website traffic. It is impossible to discuss all of them, so we focused on most commonly used like metrics: visits, pageviews, visit duration, avg time on page, bounce rate, %new visits. Such data can be analyzed in the following dimensions: the date (hour, day, month, year), location - source of visits (continent, country, city), the type, version and parameters of a web browser (IE, FF,) language etc.

B. *Data processing*

Our fuzzy SQL interpreter enables natural language expressions, so labeling module was designed and implemented. Firstly, it was necessary to define labels with the gradation of "strength" of each label (appropriate thresholds was required). Labels of the same type (e.g. short, average, tall) are combined into sets. Each label set is assigned to the attribute e.g. the attribute “time on site” can be short or long and the attribute “number of visits” can be small, average or big. The process of assigning labels to fuzzy set causes some difficulties connected to following issues [16]:

- each attribute may have different number of labels;
- each clustering process can generate different number of clusters. Presented issues were coped in the implemented algorithm.

Figure 5 presents an exemplary data distribution used in a labeling process. The naive approach is to evenly split the range of variation attribute according to the number of labels. However, it ignores the variable distribution and fuzzification process.

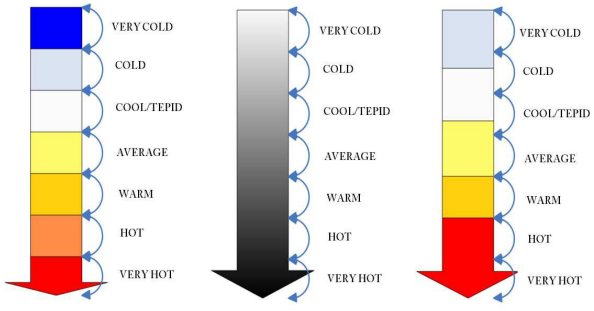


Fig 5. Exemplary data distribution used in a labeling process.

In the case of same number of clusters and assigning labels stems from the method of ordering. If number of clusters and labels are different assignment is performed by comparing the position of the centers of gravity. The last of these methods was used during creation of fuzzy query processing system.

C. *Fuzzy query processing*

In order to process the fuzzy query and generate fuzzy sets automatically, the fuzzy clustering methods, label-attribute and operator logic were implemented (Fig 6). For each distinct attribute to get most relevant results, the algorithms run several times, each time starting with new clustering parameters (different number of clusters, distance measures and start points). After the validity indexes are calculated the cluster count is adjusted on the basis of Xie-Beni and Rubio calculation for selected attribute/attributes. Next, the results of clustering method are processed by novel algorithms [13] based on membership functions in order to generated fuzzy sets with triangle and trapezoid membership functions.

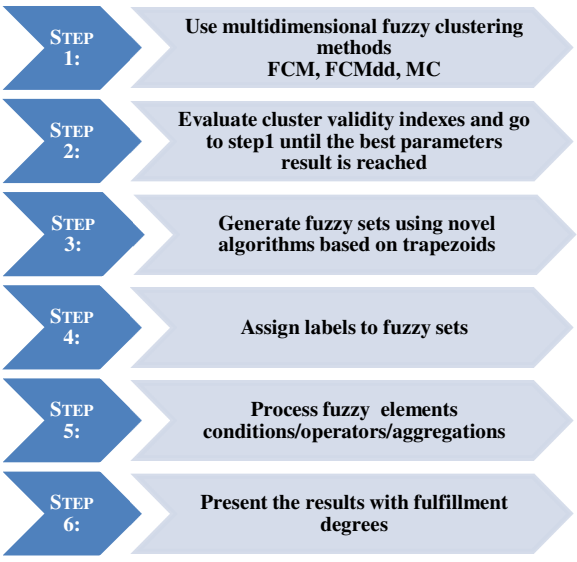


Fig 6. Multidimensional data processing.

With such prepared data the intelligent module responsible for label allocation according to the automatically generated fuzzy sets and labeling criteria, generates a set of pair-value assignments. Finally generated data collection contains fuzzy set to label matches eg. label_small » fuzzy_set_#1, label_medium » {fuzzy_set_#2 and fuzzy_set_#3} etc.

After that all the operations for current attribute based on T Norms and S Norms are processed according to query logic with respect to supplied order and operators priorities. In case of query with multiple arguments, output of processing of single arguments are joined with respect to order and operators and the final result is built from the composition of each attributes output. The mid results are combined by the use of primary keys or ROWID from primary flat SQL query. Finally the client application presents query results ordered by membership degree.

VI. EXAMPLES

SQL queries containing imprecise conditions look quite similar to traditional queries. SQL query is extended by FuzzyWhere, FuzzyHaving clauses. Figure 7 presents fuzzy query syntax, containing standard and fuzzy conditions, logical operators (or/and/not) and fuzzy aggregation conditions.

In this section, the authors present the difference between traditional SQL queries and imprecise FSQL. average number of page views during a single visit. To illustrate the possibilities of application, shown simultaneously all of available imprecise filters on all levels we can set following problem.

It is necessary to find the countries from which comes a large number of visitors, in which the average time spent on the site by users is significant.

Let's consider classical SQL syntax query solved this problem that may look like listed on figure 8. The result for query is presented on fig 9.

Similar query in FSQL language may look like listed on figure Fig 10. The result set is not restricted by crisp conditions so it contains more values Fig 11.

```
SELECT {EXPRESSIONS, AGGREGATE FUNCTIONS AGG } FROM {TABLENAME/S, VIEW/S}
WHERE {CONDITIONS}
FUZZYWHERE MINIMUMMEMBERSHIP FUZZYLABEL1(ATTRIBUTE1) AND
FUZZYLABEL2(ATTRIBUTE1)OR FUZZYLABEL1(ATTRIBUTE2) AND NOT
FUZZYLABEL3(ATTRIBUTE3)
GROUP BY {GROUP CONDITION} HAVING <CONDITIONS>
FUZZYHAVING MINIMUMMEMBERSHIP FUZZYLABEL1(AGG)
```

Fig 7. Imprecise query syntax.

```
SELECT COUNTRY, AVG (PERCENTNEWVISITS), AVG (AVGTIMEONPAGE)
FROM WEB_DATA_TST
WHERE PERCENTNEWVISITS > 90 AND AVGTIMEONPAGE > 200
GROUP BY COUNTRY
```

Fig 8. An exemplary traditional query.

COUNTRY	AVG(PERCENTNEWVISITS)	AVG(AVGTIMEONPAGE)
USA	97	225
India	96	325
Philippines	93	218
United Kingdom	92	308
Germany	92	243
Romania	90	224

Fig 9. Traditional query's results

```
SELECT COUNTRY, AVG(PERCENTNEWVISITS), AVG(AVGTIMEONPAGE)
FROM WEB_DATA_TST
FUZZYWHERE 0,5 LARGE(PERCENTNEWVISITS)AND LONG(AVGTIMEONPAGE)
GROUP BY COUNTRY
```

Fig 10. Imprecise query.

COUNTRY	AVG(PERCENTNEWVISITS)	AVG(AVGTIMEONPAGE)	SATIS
USA	97	225	0.97
India	96	325	0.96
Romania	90	224	0.95
Philippines	96	207	0.95
Germany	92	217	0.94
UK Kingdom	95	216	0.94
Brazil	94	198	0.94
Myanmar	90	196	0.93
Poland	94	196	0.93
Argentina	90	211	0.92
Canada	89	209	0.90
Tanzania	90	186	0.82

Fig 11. Imprecise query's results.

In the presented realization of query besides the definition of fuzzy labels for expressions components in FUZZYWHERE clause is possible to define the minimum cluster membership degree. This allows to control the number of elements in resulting records set. In the resulting table presents an additional column describing the degree of fulfillment of fuzzy expressions. According to the same principles we achieved the results in the case of clause FUZZYHAVING or a combination of both types of filters fuzzy.

In a situation where the traditional precise search for information is insufficient, the user has the option to take advantage of the system to define inaccurate query. Presented solution allows the user to extract information according to defined expectations and preferences. These preferences are determined at the stage of defining linguistic terms and modifications in the database. Dictionaries of labels are available during the formulation of queries. The client application enables the construction and implementation of inaccurate statements, as well as the presentation of the results of that are arranged according to the degree of fulfilment of the query.

Presented solution allows you to get fuzzy answers for the imprecise query. The fuzzy conditions can be defined in the FUZZYWHERE and FUZZYHAVING. The task is carried out in several stages, which allows

for a detailed analysis of the results of the individual phases of the query execution. Through the write access to the files mechanism, we can see both the shape of membership functions obtained as well as the result of action on individual fuzzy sets. Queries may contain aggregate functions as well as complex conditional statements and fuzzy conjunction operators and alternatives. In addition, conditions may relate to both single and multiple attributes, which is associated with the multidimensional data processing.

VII. CONCLUSION

This paper, presents a novel approach to the problem of imprecise information retrieval from database systems. SQL standard does not provide any mechanism for solving such task. Recently, fuzzy SQL languages have become a very interesting scope of research. Most of current implementations are based on strictly defined membership function and require expert knowledge about threshold degree for specific data type. This article presents an idea of gaining imprecise and incomplete information from database by novel algorithms. Most important points of the carried out research are: natural data set preparation based on real website traffic; fuzzy query processing algorithms located on database server implementation and front-end application implementation.

Main idea of the proposed algorithms is fuzzy clustering mechanisms with automatically generated fuzzy sets. All the processing is done by use of smart clustering combined with validity indexes to reach the best results. The process of generation membership function can be executed on demand, triggered by events or executed by scheduler. That feature gives opportunity to adopt to data distribution changes dynamically. In addition to this, the intelligent labeling mechanism together with own parser assigns labels defined in natural language to generated fuzzy sets. The designed system is able to execute fuzzy conditions and aggregations and can be combined with standard SQL. Currently the integration with SQL is based on Java frontend client application but in future it can be provided as an extension of standard SQL. As a summary it can be said that presented idea of fuzzy sets generator together with query parser and intelligent labeling mechanism enables retrieval of data for query written in meta-natural language (fuzzy query with smart labels).

REFERENCES

[1] Bezdek, J. (1981). Pattern Recognition with Fuzzy Objective Function Algoritms. Plenum Press.

[2] Bosc, P., Pivert, O. (1995, luty). SQLf: A Relational Database Language for Fuzzy Querying. IEEE Transactions on Fuzzy Systems, 3(1), pp. 1-17.

[3] Bosc, P., Tré, G. D., Dujmovic, J. J., HadjAli, A., Pivert, O., Ribeiro, R. (2012). On advances in soft computing applied to databases and information systems. Fuzzy Sets and Systems 196: 1-3.

[4] Buckles, B. P., Petry, F. E. (1982). A fuzzy reprezentation of data fo relational database. Fuzzy Sets and Systems(7), pp. 213-226.

[5] Buckles, B. P., Petry, F. E., Sachar, H. S. (1989). A domain calculus for fuzzy relational databases. Fuzzy Sets and Systems(29), pp. 327-340.

[6] Center for Machine Learning and Intelligent Systems, U. o. (2007). Retrieved from Machine Learning Repository: http://archive.ics.uci.edu/ml/

[7] Chang, S., Ke, J. (1979). Translation of fuzzy queries for relational database systems. IEEE Transactions on Pattern Analysis and Machine Inteligence PAMI-1, pp. 281-294.

[8] Chu, S.-C., Roddick, J. F., Pan, J. S. (2002). An Incremental Multi-Centroid, Multi-Run Sampling Scheme for k-medoids-based Algorithms – Extended Report. Knowledge Discovery and Management Laboratory; Technical Report KDM-02-003.

[9] Dembczyński, K., Przybył, D., Kalinowski, P. (2006). Retrieved from SQLf_j: http://calypso.cs.put.poznan.pl/projects/sqlf_j/pl/index.php?page=intro

[10] Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3, pp. 32-57.

[11] Dziedzic, B., Małysiak, B., Mrozek, D. (2008). Interpreter wyrażeń rozmytych stosowanych w składni języka SQL. BDAS. Ustron.

[12] Galindo, J. (n.d.). Retrieved from A Fuzzy Query Language: http://www.lcc.uma.es/~ppgg/FSQL/

[13] Pelikant, A., Kowalczyk, A. (2007). Implemntation of automatically generated membership functions based on grouping algorithms . The International Conference on "Computer as a Tool". Warsaw.

[14] Pelikant, A., Kowalczyk-Niewiadomy, A. (2009). Fuzzy queries in relational databases. System Modelling and Control.

[15] Pelikant, A., Kowalczyk-Niewiadomy, A. (2011). Algorytm etykietowania analizujący rozmyte zapytania w metajęzyku naturalnym. Bazy Danych Aplikacje i Systemy. Ustron.

[16] Rubio, E., Castillo, O., Melin, P. (2011). A new validation index for fuzzy clustering and its comparisons with other methods. Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference., (pp. 301-306). Anchorage, AK.

[17] Takahashi, Y. (1991). A fuzzy query language for relational databases. IEEE Transactions on Systems, Man and Cybernetics, 21, pp. 1576-1579.

[18] Takahashi, Y. (1993). Fuzzy database query languages and their relational completeness theorem. IEEE Transactions on Knowledge and Data Engineering, 5, pp. 122-125.

[19] [Xie, X. L., Beni, G. (1991, Aug). A validity measure for fuzzy clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on(13, 8), 841-847.

[20] Xiong, H., Zhan, G., Wu, J., Shi, Z. (2009, September). Distance Measures for Clustering Validation: Generalization and Normalization. Knowledge and Data Engineering, 21(9), 1249-1262.

[21] Yager RR, F. D. (1994). Approximate clustering via the mountain method. IEEE Trans Syst Man Cybern 24, (pp. 1279–1284).

[22] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338-353.

[23] Zamenkova, M., & Kendel, A. (1985). Implementing imprecision in information systems. Information Sciences (37(1-3)), pp. 107-141.

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 39–44

DOI: 10.15439/2015F400
ACSIS, Vol. 6

IT Infrastructure Downtime Preemption using Hybrid Machine Learning and NLP

Chiranjiv Roy, Sourov Moitra, Rashika Malhotra, Subramaniyan Srinivasan, Mainak Das

Hewlett Packard, Technology Services, GSD CSC Bangalore, India
{c.roy, sourov.moitra, rashika.malhotra, ssn, mainak.das}@hpe.com

***Abstract*—IT Infrastructure Management and server downtime have been an area of exploration by researchers and industry experts, for over a decade. Despite the research on web server downtime, system failure and fault prediction, etc., there is a void in the field of IT Infrastructure Downtime Management. Downtime in an IT Infrastructure can cause enormous financial, reputational and relationship losses for customer and vendor. Our attempt is to address this gap by developing an innovative architecture which predicts IT Infrastructure failure. We have used a hybrid approach of human-machine interaction through Big Data, Machine Learning, NLP and IR. We sourced real-time machine, operating system, application logs and unstructured case notes into an algorithm for multi-dimensional symptoms mining, using iterative deepening depth-first search, traversal to create transactions for Sequential Pattern Mining of symptoms to events. It went through multiple statistical tests and review from technology experts, to create and update a dynamic Pattern Dictionary. This dictionary is used for training unsupervised and supervised classification models of machine learning, namely SVM and Random Forrest to score and predict new logs in a real time mode. The approach is also dynamic to use unsupervised clustering methods to give directions to the technicians on future or unknown pattern of errors or fault, to constantly update the Pattern Dictionary and improve classification for new IT products.**

***General Terms*—Experimentation, Algorithms, Service Support, Technology, Research.**

***Index Terms*—IT Infrastructure Management, Big Data, Data Mining, Natural Language Processing, Information Retrieval, Decision Tree, Early Warning System, Entropy, k Nearest Neighbor Classification Algorithm, Random Forrest Classifier, Support Vector Machine, Event Clustering, Naïve Bayes Classifier.**

I. Introduction

AN UNPLANNED maintenance can be financially damaging. Marshall Institute, an asset management consultancy, reckons that a good rule of thumb is that breakdown or emergency repairs cost three times the cost of preventive, predictive or planned corrective maintenance, which still needs an end to end solution.

We, through this research introduce the concept of a holistic IT Infrastructure Downtime Management by opening a field of study to the future.

II. Related Work

Research on existing literature and industry practices reveal that many companies have recently launched services on predictive maintenance of IT equipment. Almost all services companies have some form of offering in this space. Specific book on predictive maintenance was written in early 2000 by R Keith Mobley. The literature suggests in general that we can use rule based thresholds on live monitored data of equipment and raise alerts based on rules. The literature also talks about multiple statistical and machine learning algorithms that can be used to classify new events. But the dependencies on human codified rules compel solutions to be more customized and require multiple human interventions to function. The approach presented in this paper recommends usage of multiple existing and new methods to ensure we have automated mechanisms, to build rules which then can be validated, if required. This makes the solution generic, usable across multiple types of equipment and across IT infrastructure across industries.

Regarding algorithms used most often existing solutions specify regression models to understand and forecast failures. There is also usage of binary and multiclass classification models to predict whether a failure will occur within a certain time frame. This is particularly apparent in Microsoft's Predictive Maintenance gallery model in Azure Machine learning platform. The implicit assumption in this type of usage of existing algorithms is that equipment sensor data is structured to tell us a sequence of events leading to failure. The approach in this paper suggests mechanism to work around this by identifying the pattern in sequence of events that lead to failures from raw machine data, where symptomatic events leading to final failure are not explicit in nature.

Another important area which was very apparent from research surveys but mostly missing in existing industry implementation of predictive maintenance is Gartner and IDC's 80-20 rule: 80% of all downtime is due to people and process issues and 20% due to technology issues. Over dependence on machine sensors constraints the model to look and predict a machine's performance. This leads to an oversight of people and process issues. Our work in this paper highlights that with generic pattern identification and subsequent classification based on validated patterns can help capture people and process issues as well.

Thus the paper tries to extend existing related work into generic predictive maintenance solution which is applicable to an environment with varied IT infrastructure. People and process issues are also predicted in this generic approach.

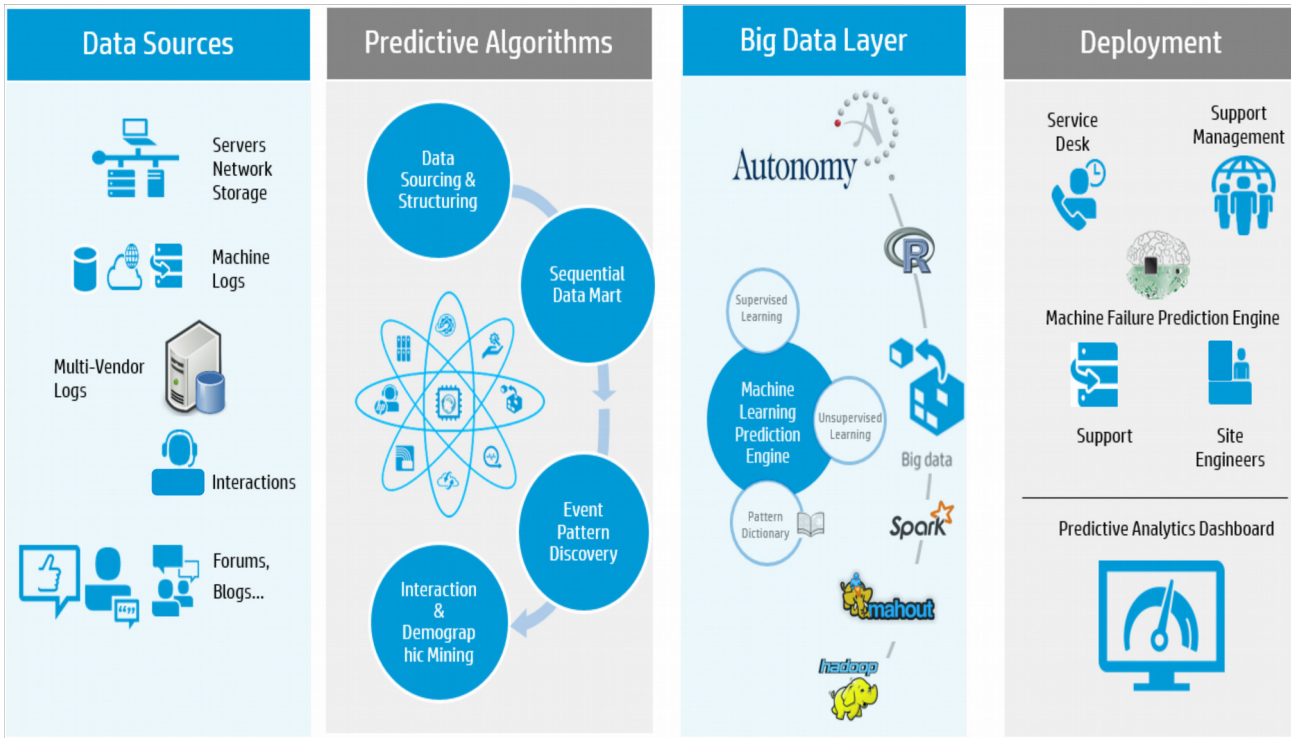


Fig 1. High Level Diagram

III. High Level Design

Figure 1 gives an overview of our approach

IV. Data Sourcing

Machine performance data is stored as Log, and is the main resource to understand machine behavior. This can be collected over network using programs written in different languages. An Enterprise IT environment uses different types of equipment based on the usage. The challenges are:

a) Variety of the Product: Servers, Storage, Networking, power backups etc.
b) Multi manufacturer: Same types of the product from different manufacturer placed in one environment.
c) Types of machine Data: every machine has its own style of providing data. Some of them are plain text, some of them are encrypted or as a binary file with different types of data structure

The challenge is combining all of the data to one readable and usable format for the algorithm to understand and apply models for study and usage. Above all, the challenges of doing data mining and modeling on such large data are many. Statistically surmising, continuous and discrete data together has always been a challenge. In a huge dataset with multiple sources, with multiple pieces of information from independent sources, it is very difficult to infer something based on looking at just a few pieces of information; as different permutations of multiple pieces of information could lead to various conflicting events. The data is also dynamic, as logs are continuously fed. On top of that, the inflow of information is fast paced, huge and requires immediate attention to trigger corrective actions on time. So unless and until the snapshot of the data is captured and analyzed with due consideration to timeline, it can lead to incoherent inferences.

The challenges don't just end there. The lack of centralized data base with multiple systems handling different type of data for different domains and with lack of proper governance across them, a lot of data problems like duplicity, missing data, wrong data, improper time tracking etc. also arise. The real challenge comes in tracing whether the data is correct or not. From analysis perspective, the "descriptions" for each line of log are captured as unstructured data; which if analyzed accurately using advanced text mining algorithms, along with Machine Learning Supervised methods, in an Hybrid Learning Method, can give great insights into the Downtime Management, and if missed can lead to erroneous conclusions.

V. Discovering Patterns–The Supervised Way

A. Developing Symptoms Database

Let us elucidate the need for self-learning Transactional Table and how we have conceptualized it. Each source of data becomes very dynamic post decrypting in terms of flow of events and defining relationships to me mined. We studied and found that an enterprise data of such a large scale does not have a standardized way to collapsing all the sources in a robust format for further study. The generic approaches as illustrated in [20][21][22][23][24][29] cannot be directly applied because of diverse systems resulting in complex behavior as seen in the logs. A sample as shown in table (1) illustrates the complexity in back and forth rapidly changing dimensions of the Symptoms to Events generation. The final result may also be skewed towards either a complete failure

TABLE I.
SYMPTOMS TABLE

Symptom 1 (S_1)	Symptom 2 (S_2)	Symptom N (S_n)	Event
S_{11}	S_{13}	S_{52}	E_1
S_{21}	S_{44}	S_{57}	E_1
S_{17}	S_{31}	S_{66}	E_2

of the system or a symptom or an event, hence complexity multiplies itself.

B. Event Pattern Discovery

Pattern mining or discovery is a holistic concept and has been applied in Marketing Science [25][26] and other research fields [27] for decades now. Analyzing symptoms from billions of rows into meaningful patterns in rapidly changing environments like IT Infrastructure. We had to take a non-traditional way to approach this problem. We first created a framework of "Pattern Dictionary" which works in two ways. We developed a path based combined length and breadth search function as illustrated in papers in bits and pieces in studies [15][16][17][18][19]. The result of this exercise was to find the best possible pattern using multi-level search and precisely try to churn the duplicates at each hierarchical level of system, server or entire environment and find best known patterns.

The second way was to apply Constraint based Sequential Pattern Mining [22] [23] [24], to study and find the unknown patterns using pattern growth method for frequent pattern mining. The sequence of the data was determined by the transactions by finding events occurring at regular intervals and then finding difference between two similar events.

We then applied a visual Self Organizing Map on the rules to determine the rules generated with the in-house function and sequential pattern miner (example as shown in Figure 2). The resultant patterns are clubbed together and gets into the cleansing process.

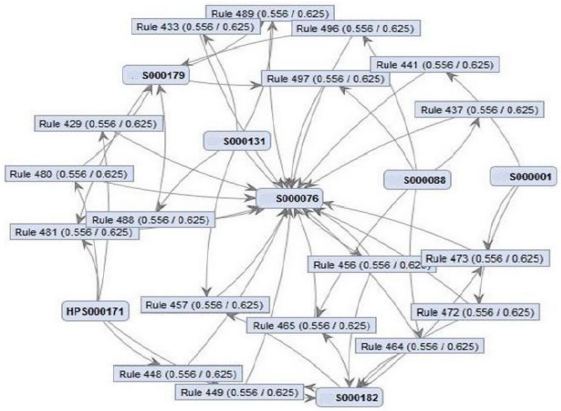


Fig 2. Rule Network using Self Organizing Map

A sequence $\alpha = (X1 \bullet\bullet\bullet Xl)$ is an ordered list of itemsets or symptoms in our case. An itemset Xi $(1 \le i \le l)$ in a sequence is called a transaction, a term originated from analyzing customers' shopping sequences in a transaction database.

The number of transactions in a sequence is called the 'length of the sequence'. A sequence with length l is called an l-sequence. For an l-sequence α, we have $len(\alpha) = l$. Furthermore, the i-th itemset is denoted by α[i]. An item can occur at most once in an itemset, but can occur multiple times in various itemsets in a sequence.

A sequence $\alpha = (X1 \ldots X_n)$ is called a subsequence of another sequence $\beta = (Y1 \ldots Y_m)$ $(n \le m)$, and β a super-sequence of α, denoted by α !:; β, if there exist integers $1 \le i1 < \ldots < in \le m$ such that $X1 \subseteq Yi1, \ldots, X_n \subseteq Y_{in}$.

A sequence database is a set of 2-tuples (sid, α), where sid is a sequence-id and α a sequence. A tuple (sid, α) in a sequence database SDB is said to contain a sequence γ if γ is a subsequence of α. The number of tuples in a sequence database SDB containing sequence γ is called the support of γ , denoted by sup(γ).

Given a positive integer min_sup as the support threshold, a sequence γ is a sequential pattern in sequence database "Patterns Database" if $sup(\gamma) \ge min\_sup$. The sequential pattern mining problem is to find the complete set of sequential patterns with respect to a given sequence database SDB and a support threshold min_sup.

The first principle that we apply here is applying a weightage based on Pearson's chi-square test, also known as the chi-square goodness-of-fit test or chi-square test for independence [32]. This gives an advantage over and above choosing the rules based on lift and support. The result from association rules P_1 passes on to the Human Machine Interface (HMI) for next level of check.

C. Human Machine Interface for Pattern Validation

Patterns from the mining activity as illustrated above is treated in the algorithm using a new and robust mechanism of throwing the first patterns P_1 to be matched with the generic list of error databases existent with all enterprise support organizations. The resultant is Pattern List 2 (P_2) which is then passed to the experts and technicians to validate the patterns with their experiences through a web portal in cloud to capture their validations using voting methodology of event occurrence. The weight of voting on specific pattern clears its path for the self-learning and ever evolving "Pattern Dictionary". This eliminates bias from statistical and industrial application perspective as shown in Table 2. A java based portal ($\mathbf{J_p}$) is used to show the entire frame and the data is stored in the HP Vertica data center.

TABLE II.
SYMPTOMS TABLE

Pattern	Event	Lift	Vote	P_{Votes}
$S_{11} > S_{43} > S_{91}$	E_{11}	0.69	Yes	39%
$S_{43} > S_{56} > S_{45} > S_{67}$	E_{23}	0.35	No	77%

VI. Predicting The Next Event–Supervised Machine Learning Approach

Now that the "Pattern Dictionary" is ready with the patterns from the exercise done in Step 5, it is then needed to predict the next event occurrence. Decision Trees, [13] classically try to address this problem at a smaller scale and dimension (including limited classes). But at a big data level, we thought of applying Machine Learning in a different way on results from Association rule applied in some studies[1] [2][3][4][5][6][7][8][9]. We have tried to study some machine learning models which is one the way of being implemented using Python and Spark.

A. Application of Machine Learning models with supervision

Every generation of machines may have a different behavior and may generate varied class of events. Using the logic of determining the best fit model from historical data may collapse when applied to real time. Hence we approached this problem by providing the flexibility of choosing the best fit model for a particular one real time, using a Java based portal ($\mathbf{J_p}$).

We applied few machine learning models like kNN, Naïve Bayes, Support Vector Machine and Random Forest which can be scaled to Big Data framework at run time using Spark or Mahout (development in progress).

The above approach is quite intuitive in nature to capture complexities of the framework and give an advantage to the user or technician to apply the best fit model on real time. The first cut of results are shown in Table 3 above which tells us that the Sigmoid Kernel function used for SVM tends to give a better result but not far behind is the Gradient boosting model with 52 % classification accuracy. As there is no existing model to test with, this testing model would be taken as reference for further evaluation.

B. Determining unknown events through Text Clustering and NLP

In the process of applying supervised mechanisms to predict next events we found that a significant percentage of events could not be classified or be predicted due to historical novelty, new machines developed or implemented, software updated, patch released etc. These may create a huge pile of symptoms which could not even be mined or mapped by human intelligence.

We propose a dynamic flexibility to the algorithm which takes the key aspects of Text Mining, Natural Language processing and clustering documents.

The first challenge was to put the millions of unstructured and structured data generated through server monitoring and collected through case logs, into a centralized data repository, on top of which an analytical system can be built. It was essentially a Big Data challenge, for which a HP Vertica Platform was used, to address it. Once the data was available, the next challenge was to apply machine learning on this Big Data. R MapReduce code came handy here to generate Document Term Matrix and Mahout was used for Clustering and Classification [31][32].

A clear roadmap was laid and followed to develop a Hybrid Pattern Mining and Unknown Issue Identification System. The incidents (symptoms) were divided into two sets: one where the affected items were mentioned and the other where the affected items were not mentioned. The incidents where the affected items were mentioned were further divided into the training and validation data sets. The training set was used to derive the affected items for incidents where it was not mentioned using machine learning (Pattern Min-

Table III.
Sample Results from Machine Learning Models

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Misclassification Rate	Squared Error	Actual: Misclassification Rate
Y_{Event_A}	SVM	SVM	SVM-d	Final_Event	0.13582	0.13509	0.13582
	SVM3	SVM3	SVM-FQP	Final_Event	0.13582	0.13509	0.13582
	HPNNA2	HPNNA2	Select NN	Final_Event	0.13583	0.11834	0.13583
	HPNNA	HPNNA	HP Neural	Final_Event	0.13583	0.11988	0.13583
	LARS	LARS	LARS-d	Final_Event	0.13666	0.11319	0.13666
	SVM2	SVM2	SVM-LSVM	Final_Event	0.13674	0.14751	0.13674
	Reg	Reg	Regression-d	Final_Event	0.13682	0.11329	0.13682
	SVM4	SVM4	SVM-Sigmoid Kernel	Final_Event	0.54437	0.26851	0.54437
	Boost2	Boost2	Gradient Boosting - hash1	Final_Event	0.52273	0.52273	0.52273
	Boost	Boost	Gradient Boosting -d	Final_Event	0.13582	0.11712	0.13582

ing). The validation data set was used for in-sample validation. Once the affected items were known for all the incidents, the next step was to group those millions of independent incidents into logically related units which can be used to design an Incident Routing Mechanism. This mechanism ensures that a randomly reported incident is correctly allocated to its designate Level II tech support team on time, for accurate and timely resolution of the issue.

The Pattern Mining was done independently on a sample of top 5 clients by incidents/issues logged by call volume collected the free form textual inputs given by multi-level support representatives and server logs generated for those accounts connecting logs with R-Vertica Connect. First data cleansing was done including removal of punctuations, white spaces, numbers, stop-words, etc. The stop-word removal was a two-step process: first a standard corpus based and second specific to each account to take out many English word which are required for NLP but not needed in this case. The account based list was created. Next POS (parts-of-speech) tagging was done and only nouns and verbs with alphanumeric keys were kept. The alphanumeric keys was identified using N-gram taxonomy created and verified by experts. Finally, after all the cleansing, the DTM (Document Term Matrix) was created to link the term to the affected items. The DTM was shared with the engineers or experts to further refine it. The industry best practice of 10-fold cross validation technique was then used to find the relation between the unstructured incidents listed in the form free text and the affected items. The final DTM list was divided in training (90%) and validation (10%) sets, randomly 10 times. The kNN (k-nearest-neighbor) classification algorithm was run every time on the training set and the validation set was used to find the misclassification rate (error). The desired value of k (from 1to 10) was selected based on the classification which gave the least misclassification rate (error). After the in-sample validation was done and we got both the right value of k and the relation between the unstructured incidents and the affected items; the findings were used to derive the affected items of those incidents where the affected items were not listed using kNN classification again. The output was a complete list of all the incidents with affected items.

The event determining mechanism was more challenging as there was no predefined/business logic behind grouping of similar incidents together and it was purely based on un-supervised machine learning algorithm. The k-means clustering algorithm of Mahout was chosen for the purpose. The biggest challenge here was to find the correct value of k. Silhouette approach was used to find the right value of k. In this method the items within each cluster are evaluated for their average dissimilarity with all the other items in the cluster. The measure of dissimilarity was selected as Euclidian Distance, say a(i). The item is similarly evaluated for its average dissimilarity with any other cluster to which it does not belong, say b(i). The cluster with the lowest average dissimilarity is said to be the 'neighboring cluster'. If $a(i) < b(i)$, then it can be concluded that the item belongs to the right cluster, else the clusters are revised. Thus the Silhouette approach is used to determine the value of k based on how tightly grouped all the items in a cluster are.

The choice of Silhouette method posed another challenge as it requires the data to be in the form of a non-negative square matrix to be run on Mahout. SVD (Singular Value Decomposition) technique was used to address this issue [28]. The DTM for the entire data set was created the same way as specified in the pattern mining section above, but for the use of Mahout instead of R. The DTM was transformed using SVD which not only resulted in a square matrix but also brought in dimensionality reduction; reducing the space complexity of the Silhouette algorithm which becomes highly space complex for Big Data. Once the logical grouping of incidents based on terms identified from the unstructured data in the form of comments for each incident was completed, the routing mechanism to decide which should be routed to which support team was very easy to develop.

These patterns are displayed online using the portal to map closely related events and results are saved in the Pattern dictionary.

C. *Bridging the Supervised and Unsupervised approach*

The "Pattern Dictionary" acts as the stakeholder for collecting information from both the supervised and unsupervised methods. These patterns may not be conclusive but are evolutionary in nature which develops as time grows and studies each system working as a perfect interface between Human and Machine.

The approach collects all the necessary pattern into the pattern dictionary and then applies multiple models to be tested as we do not know which model will fit better in what kind of situation.

Historical knowledge and Topics gained from the NLP approach makes this concept unique and a perfect bridge between Supervised and Unsupervised Methods.

VII. Conclusion

This research with Predictive Analytics in the center, is an attempt to preempt failures in a customer's IT environment using an algorithm with an approach starting with Data sourcing & structuring, transforming it for unique pattern discovery using the theory of Ripple effect from Symptoms, to Events for any IT infrastructure. This also uses external information like weather forecasting, blogs, customer surveys, social media to strengthen the hybrid machine learning algorithm. This is a scalable Big Data Solution, which predicts real time faults and downtime.

The benefit of this is that solutions are many including - prevention of unplanned downtime, maximizing the value of IT investments by reducing the cost, improving operational Efficiency, enabling Agility and Innovation, improving IT stability – less complexity and risk.

With this novel approach of using machine learning in a real world problem with multiple complexities, it tends to open a new field of study with integrated machines, servers, equipment and its impact on an overall IT environment.

VIII. Future Work

Due to the rapidly changing dynamic of the IT industry and the challenges faced with IT Infrastructure Management

as well as IT Infrastructure Maintenance, which is a billion of dollar industry in itself, it is the need of the hour to conduct research and support it. A novel amalgamative concept is what we have tried and briefly presented here.

We are still in the phase of learning as there is no research done so far on this field, hence the next phase is to try and first complete the solution in a big data scale after multiple testing cycles to make it robust enough for developing it as a solution for the first version.

The area we haven't explored much is the relationship mining of customer interactions and machine. Also optimizing it using Neural Network or Genetic Optimization techniques.

References

[1] Aggarwal, Charu C., Yu, Philip C. 2001. Outlier Detection for High Dimensional Data, ACM SIGMOD

[2] Ghose, Udayan., Rai, C.S., Singh, Yogesh. 2010. On Multiplicative Entropy and Information gain in Large Data Sets, International Journal of Engineering Science and Technology, 187-193.

[3] Han, Jiawei., Kamber, Micheline., Pei, Jian. 2011. Data mining: Concepts and Techniques, 561-562, Morgan Kaufmann.

[4] Hodge, Victoria J., Austin, Jim. 2004. A Survey of Outlier Detection Methodologies, In: Artificial Intelligence Review, 85-126, Kluwer Academic Publishers, Netherlands.

[5] Knorr, Edwin M., Ng Raymond T. 1998. Algorithms for Mining Distance-Based Outliers in Large Datasets, VLDB Conference.

[6] Minka, Thomas P. 2003. A comparison of numerical optimizers for logistic regression.

[7] Pawling, Alec., Chawla, Nitesh V., Chaudhary, Amitabh. 2005. Computing Information Gain in Data Streams, Temporal Data Mining Workshop.

[8] Pliner, Vadim. 2004. A SAS® Macro for Naïve Bayes Classification.

[9] Pokrajac, Dragoljub., Lazarevic, Aleksandar., Latecki, Longin Jan. 2007. Incremental Local Outlier Detection for Data Streams, IEE Symposium on Computational Intelligence and Data Mining (CIDM).

[10] Rokach, Lior, Maimon, Oded. 2010. Decision Trees. In: Data Mining and Knowledge Discovery Handbook, 165-192, Springer.

[11] Sahami, Mehran.1996. Learning Limited Dependence Bayesian Classifiers.

[12] Tan, Pang-Ning., Stienbach, Michael., Kumar, Vipin. 2007. Introduction to Data Mining, 139-20, Pearson.

[13] Agrawal, R., Amielinski, T., and Swami, A. (1993). Mining association rule between sets of items in large databases. In Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, DC, May 26-28.

[14] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rule. Proceedings of the 20th International Conference on Very Large Data Bases. pp. 487 – 499.

[15] Antonie, M., Zaïane, O. R., Coman, A. (2003). Associative Classifiers for Medical Images. Lecture Notes in Artificial Intelligence 2797, Mining Multimedia and Complex Data, pp 68-83, Springer-Verlag.

[16] Blackmore, K. and Bossomaier, T. J. (2003). Comparison of See5 and J48.PART Algorithms for Missing Persons Profiling. Technical report. Charles Sturt University, Australia.

[17] Brin, S., Motwani, R., Ullman, J., Tsur, S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data.

[18] Cendrowska, J. (1987). MODEL: An algorithm for inducing modular rules. International Journal of Man-Machine Studies. Vol.27, No.4, pp.349-370.

[19] Cohen, W. W. (1995). Fast effective rule induction. In the Proceeding of the 12 th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, pp. 115-123.

[20] Cohen, W. W. (1993). Efficient pruning methods for separate-and-conquer rule learning systems. In the proceeding of the 13th International Joint Conference on AI, Chambry, France.

[21] Cowling, P. and Chakhlevitch, K. (2003). Hyperheuristics for Managing a Large Collection of Low Level Heuristics to Schedule Personnel. Proceeding of 2003 IEEE conference on Evolutionary Computation, Canberra, Australia, 8-12 Dec 2003.

[22] Dong, G., Li, J. (1999). Efficient mining of frequent patterns: Discovering trends and differences. In Proceeding of SIGKDD 1999, San Diego, California.

[23] Chris Buckley and Alan F. Lewit, Optimizations of inverted vector searches, SIGIR '85, Pages 97-110, 1985..

[24] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1996). Advances in knowledge discovery and data mining, MIT Press.

[25] Zaki, M. J., Parthasarathy, S., Ogihara, M., and Li, W. (1997). New algorithms for fast discovery of association rules. 3rd KDD Conference, pp. 283-286, August 1997.

[26] Charu C. Aggarwal, Stephen C. Gates and Philip S. Yu, On the merits of building categorization systems by supervised clustering, Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 352 – 356, 1999.

[27] Paul Bradley and Usama Fayyad, Refining Initial Points for K-Means Clustering, Proceedings of the Fifteenth International Conference on Machine Learning ICML98, Pages 91-99. Morgan Kaufmann, San Francisco, 1998

[28] Alvarez, Sergio A. Technical Report BC-CS-2003-01, July 2003. Chi-squared computation for association rules: preliminary results

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 45–52
DOI: 10.15439/2015F153
ACSIS, Vol. 6

Implementation of Decision Support System on m/f Wolin.

Piotr Wołejsza
Maritime University of Szczecin
Wały Chrobrego Str. 1-2, 70-500
Szczecin, Poland
Email: piotr@am.szczecin.pl

***Abstract* — The known navigational systems in use and methods of navigational decision support perform information functions and as such are helpful in the process of safe conduct of a vessel. However, none of these known systems provides a navigator with ready solutions of collision situations taking account of all the vessels in the proximity of own ship, where the Collision Regulations apply. This paper presents verification results of NAVDEC – new Navigational Decision Supporting System created by research team from Szczecin Maritime University both for ocean going ships and pleasure crafts. Successful verification was carried out in real condition on board Motor Ferry Wolin (m/f Wolin), which belongs to shipowner UnityLine.**

I. Introduction

THE competitive position of maritime transport compared to the other transport modes leads to a continuous increase in the carriage of goods by sea, which entails higher traffic intensity, vessel tonnages and speeds. This, in turn, adversely affects the safety of people, ships, cargo and marine environment. To enhance navigational safety, efficiency and competitiveness of transport services in maritime trade, both ships' and land-based vessel traffic centres' equipment and systems are constantly being upgraded. Such facilities perform mainly information functions and in this respect they support the process of safe ship conduct. However, the amount of information available on the ship has been on the rise while the technical systems have become more complex. For these reasons both information management and the resultant decision making are difficult, e.g. emergency situations may go beyond decision-maker's abilities.

A review of maritime court decisions indicates that human errors are one of the major causes of marine accidents. Elimination or reduction of human errors, which would provide for possibly high safety level, can be achieved only by equipping ships with tools that, apart from information functions, will work out solutions to collision situations accompanied by adequate comments. None of the systems known to date is capable of performing such functions. Therefore, decision support is rather restricted, and, consequently, collisions sometimes are not avoided. A higher level of navigational safety gained through the introduction of the system performing the new functionalities will reduce the risk of marine accidents. This will bring the following advantages:

- social benefits due to lower rate of personnel injuries and loss of life on sea-going ships,
- material benefits due to lower loss of cargo, less damage to ships or sinkings,
- marine environment protection and prevention of ecological disasters that occur as a consequence of collision of ships carrying dangerous goods.

The navigational decision support system NAVDEC is the first navigational tool worldwide that performs information functions as well as those typical of decision support systems. Its innovative functionalities, significantly extending the performance of devices generally carried by ships, have now a status of patent applications filed at home and internationally.

II. Test environment

M/F Wolin is a train, car and passenger ferry owned by the Unity Line. She was built in 1986. It has been in Unity Line colours since 2007. It is the longest ferry on Świnoujście - Trelleborg route.

Fig. 1 M/F Wolin [1]

The author wants to thank for the financial support to Ministry of Science and Higher Education.

Call sign	C6WN4
Length	188,9 m
Breadth	23,1 m
Draught	5,9 m
Maximum speed	18 knots
Crew	37
Passengers capacity	370
Cabins	70
Beds	240

First test installation took place 14-15 of September 2014. During the journey to/from Trelleborg few deficiencies were observed. Corrections in the source code were done and from 16th of December 2014 NAVDEC is in continuous use on m/f Wolin.

NAVDEC is installed on the portable computer Hewlett Packard ProBook 6555b (processor ADM Athlon II P340 Dual Core 2.2 GHz, hard drive 256 GB, 2 GB RAM) with 32 bit Windows 7 together with C-MAP Professional + chart license. Computer is connected to ship's system via two cables:

1. AIS (Automatic Identification System) Pilot Plug to RS 232,
2. ARPA (Automatic Radar Plotting Aids) to RS 232

cargo and/or environmental damage), is more than one million USD, while the average sum of hull and machinery damages paid by insurance companies is of two billion USD yearly [2]. The European Quality Shipping Information System database (www.equasis.org) quantifies merchant vessels at over 77 thousand worldwide, of which around 2.5% come into collision every year, while the Helsinki Commission (HELCOM), which monitors the Baltic Sea, cites collisions as the main type of accidents, accounting for 38% of the total in that area [4].

The Swedish Club report concludes that the majority of these collisions could have been prevented by following safety protocols, but the fact is that sea-going vessels lack dedicated support systems to address potential collision situations. This is the purpose of NAVDEC, a navigational decision system designed to support vessel officers in efficient collision avoidance following COLREGs rules. The need for such systems has been confirmed by the International Maritime Organization forum developing e-navigation strategies [5] and by key stakeholders involved in the development of the NAVDEC prototype currently pilot testing in real work environment (TRL 7).

NAVDEC uses the navigation system data to provide fast and accurate options to the OOW, including the most important variables to be considered:

Fig. 2 NAVDEC in action: real screenshot from the NAVDEC pilot test at sea on the **Dar Mlodziezy**, showing main components.

III. NAVDEC

Marine accidents represent a major risk for personnel, cargo, vessels and the environment. According to marine mutual insurer The Swedish Club, one of the main causes of vessel collisions is that the Officer Of the Watch (OOW) did not follow International Regulations for Preventing Collisions at Sea (COLREGs) or their company's Safety Management System [2]. Marine insurance statistics have shown that human error is a major contributing factor in about 60% of shipping accidents, with other research suggesting that this figure significantly increases in the case of collisions and groundings [3].

The average cost of ship collision, taking account hull repair cost alone (excluding the costs of medical care, lost

1. Classification of encounter situation according to COLREGs (**"crossing situation", "head on situation"** or **"overtaking"**) and which vessel is **"stand on"** (has right of way) and **"give way"** (must let the other pass)
2. Optimal course to avoid collision
3. Target data
4. Solutions for selected target
5. Own vessel data: destination, distance to go, ETA, etc.
6. Solutions how to pass all targets at predetermined distance from our vessel.

When the own vessel is "give way" in relation to at least one target, NAVDEC displays a compass rosette with solutions (6), with red sectors indicating collision risk and

yellow sectors indicating safe courses in which the vessel will pass other targets on predetermined CPA or larger distance. Below the rose is the optimal course requiring smallest deviation from current course: a green arrow indicates starboard turn, red arrow for port turn.

Navigational systems installed on vessels are information systems which acquire, process, gather and display information to the ship's navigator, who makes decisions after analysing the data and assessing the situation. Rough weather conditions, heavy traffic, stress or fatigue may provoke errors in situation assessment and lead to a wrong decision. The implementation of decision support will help reduce the number of such errors and enhance the safety and efficiency of maritime transport, while also leveraging a necessary technological component of future unmanned e-navigation vessels. Moreover, the system optimizes the anti-collision manoeuvres to reduce fuel consumption, which is the main component of the cost of transport. [6]

In a 2014 report by the NCSR IMO Sub-Committee in London, the e-navigation group of the International Maritime Organization stated *"[...] It is important to recognize that further e-navigation development will be a continuous process following user needs for additional functionalities of existing and possible future systems (e.g. implementation of onboard and/or ashore navigational decision support systems)."*[5]

One of the ways to reduce the number and consequences of marine accidents is the application of shipborne navigational systems that, apart from information, perform decision support functions: automatically generate suggested solutions to collision situations, leaving the choice to the navigator. The lack of such systems on the market creates opportunities for companies dealing with their production and implementation.

NAVDEC is a real time system handled by the navigator that complements the navigational equipment of the ship. The system observes its ship and the environment and records information on the present navigational situation. On this basis the system identifies and assesses the navigational situation (processing) and works out solutions (decisions) assuring safe navigation.

For the system to function correctly it must cooperate with standard equipment and systems installed on board (often used on leisure craft as well) such as: log, gyrocompass, ARPA, GNSS (Global Navigational Satellite System), AIS (Automatic Identification System), ENC (Electronic Navigational Chart) and sources of current navigational data. NAVDEC performs information functions – on one screen it presents bathymetric data from an electronic chart, an image of surface situation from a tracking radar, positional information from the AIS and GNSS receivers. Finally, it determines and presents to the navigator movement parameters of targets in vicinity. [7]

The navigational decision support system NAVDEC is the first navigational tool worldwide that performs information functions as well as those typical of decision support systems. NAVDEC goes beyond the current functions of information systems such as ECDIS (Electronic Chart Display and Information System) and ARPA (Automatic Radar Plotting Aids) by offering the following advanced functions:

- fusion and integration of navigational data received from shipboard devices
- analysis and assessment of situation taking into consideration the Collision Regulations in force
- automatic determination of solutions to collision situations by using dedicated computational algorithms
- explanation of the present navigational situation making use of a navigational knowledge base (collision regulations, principles of good sea practice, criteria of navigational situation analysis and assessment actually used by expert navigators)
- justification of the recommended manoeuvre. [8]

Compared to the ARPA system, presently used on ships for calculating ship encounter parameters and working out an anti-collision manoeuvre, the developed NAVDEC system has the following advantages:

- account for the Collision Regulations for good and restricted visibility,
- generates a manoeuvre in relation to other ships, also those located in the radar blind sector,
- operator is immediately notified of a manoeuvre commenced by another ship (target) thanks to information about rate of turn of the target, and the system needs a few seconds to calculate encounter parameters, while ARPA, according to IMO's test situations, needs three minutes for this action,
- more accurately calculates encounter parameters, as it takes into account ship's dimensions, and uses GPS for position determination by special algorithms executing data fusion,
- takes into account the sizes of ships while planning an anti-collision manoeuvre,
- calculates such new courses and speeds of own ship that passing other targets is possible maintaining the predetermined closest point of approach (CPA). [9]

IV. DECISION SUPPORT SYSTEM

Developed at the Maritime University of Szczecin NAVDEC system is a navigation tool that performs alongside providing information typical tasks for decision support systems. NAVDEC is an important complement to navigational equipment of the ship. Is a real-time system operated by the navigator. Its proper functioning requires interaction with devices and systems on the ship. The standard configuration of the ship include: log, gyrocompass, radar, echo sounder, ARPA, GNSS (Global Navigational Satellite System), such as GPS (Global Positioning System)

or DGPS (Differential Global Positioning System). In addition, AIS, ECDIS, GNSS [10]. In the version being developed following sources of information are in use: log, gyrocompass, radar / ARPA, GPS and DGPS, AIS and ENC (Fig.3).

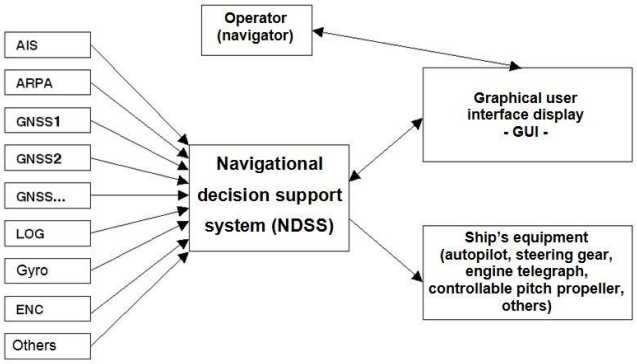


Fig. 3. Data sources for decision supporting system [9]

The system's structure (Fig.4) has been prepared in such a way so as to make possible the simultaneous performance of tasks bound with supporting decision-making processes in the conduct of a ship.

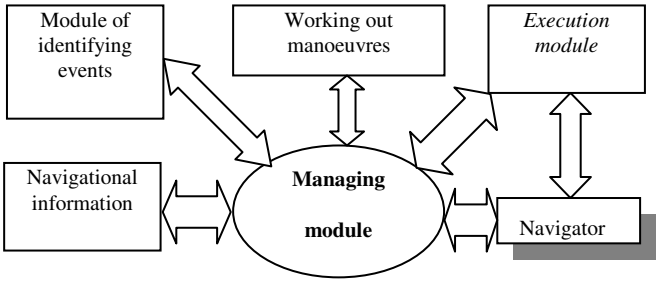


Fig. 4. Multi-agent system structure [6]

The system's functioning is based on an algorithm consisting of eleven procedures:

- determining risk of collision (***PROCEDURE 1***),
- determining phase of encounter (***PROCEDURE 2***),
- determining right-of-way in relation to extraneous vessels (***PROCEDURE 3***),
- calculating changes of course and speed leading to mutual passing on a preset CPA (Closest Point of Approach) (***PROCEDURE 4***),
- working out a manoeuvre (***PROCEDURE 5***),
- last-moment manoeuvre (***PROCEDURE 6***),
- admissible intervals of course and speed alterations (***PROCEDURE 7***),
- decreasing the assumed CPA (***PROCEDURE 8***),
- manoeuvre in relation to vessel with smallest TCPA (Time to Closest Point of Approach) (***PROCEDURE 9***),
- optimisation of manoeuvre (***PROCEDURE 10***),
- performing the manoeuvre (***PROCEDURE 11***).

Basic criteria for the assessment of the navigational distance are Closest Point of Approach (CPA) and Time to Closest Point of Approach (TCPA). They are commonly used in Automatic Radar Plotting Aids (ARPA).

$$CPA = \frac{X_{wz}\, VY_{wz} - Y_{wz}\, VX_{wz}}{V_w} \qquad (1)$$

$$TCPA = \frac{X_{wz}\, VY_{wz} - Y_{wz}\, VX_{wz}}{V_w^2} \qquad (2)$$

where:

VX_{wz}, VY_{wz} – relative speed vector components,
X_{wz}, Y_{wz} – distance between vessels counted along x and y axes, respectively,
V_w – relative speed.

Determination of the ship's own course for the passing of an object at a given distance is possible depending on the analytical [11]:

$$\operatorname{tg}\frac{\psi}{2} = \frac{A_{CPA}\, V \pm \sqrt{(A_{CPA}^2 + 1)\, V^2 - B_{CPA}}}{B_{CPA} - V} \qquad (3)$$

$$A_{DCPA} = \frac{X_{wz}\, Y_{wz} \pm CPA\sqrt{D^2 - CPA^2}}{X_{wz} - CPA^2} \qquad (4)$$

$$B_{DCPA} = A_{CPA}\, V_x - V_y \qquad (5)$$

where:

V – own ship speed,
Xwz, Ywz – distance between vessels counted along x and y axes respectively,
V_x, V_y – components of the velocity vector of own ship,
D – distance between vessels,
ψ – new course which enables to pass other targets on presumed CPA.

In a similar way it is possible to determine the speed of own ship, which enables to pass other targets on presumed CPA.

Basing on above equations following source code was developed. The courses leading to pass at presumed distance (*Safe_Courses* procedure). The courses are calculated on the basis of [5] for each pair: the own ship (number 1) and the target ship *i* (for *i* = 2 to *n*, where *n* is the number of target ship).

Input data:

- position (*x*1, *y*1), speed (*V*1) and course over ground (*KDd*1) of the own ship,
- position (*xi*, *yi*), speed (*Vi*) and course over ground (*KDdi*) of target,
- CPA – safe passing distance set up by navigator

Output data:

< gamma$_i$1, gamma$_i$2 >, <gamma$_i$3, gamma$_i$4> - sectors of safe courses for pair: the own ship and the target ship number *i*.

Safe_Courses(i):

```
{
xwz=xi-x1; ywz=yi-y1; vxwz=vxi-vx1;
vywz=vyi-vy1;
vw=sqrt(vxwz*vxwz+vywz*vywz);
D=sqrt((xwz*xwz+ywz*ywz)^2);
Adcpa1=(xwz*ywz + CPA * sqrt(D^2-(CPA)^2))
/(xwz*xwz - (CPA)^2);
Adcpa2=(xwz*ywz - CPA * sqrt(D^2-(CPA)^2))
/(xwz*xwz -(CPA)^2);
vxi=Vi*sin(KDd2); vyi=Vi*cos(KDdi);
Bdcpa1=Adcpa1*vxi-vyi;Bdcpa2=Adcpa2*vxi-
vyi;
gamma_i1=2*atan((Adcpa1*V1+sqrt((Adcpa1*Ad
cpa1+1)*V1*V1-Bdcpa1*Bdcpa1))/(Bdcpa1-
V1))*180/pi;
gamma_i2=2*atan((Adcpa1*V1-
sqrt((Adcpa1*Adcpa1+1)*V1*V1-
Bdcpa1*Bdcpa1))/(Bdcpa1-V1))*180/pi;
gamma_i3=2*atan((Adcpa2*V1+sqrt((Adcpa2*Ad
cpa2+1)*V1*V1-Bdcpa2*Bdcpa2))/(Bdcpa2-
V1))*180/pi;
gamma_i4=2*atan((Adcpa2*V1-
sqrt((Adcpa2*Adcpa2+1)*V1*V1-
Bdcpa2*Bdcpa2))/(Bdcpa2-V1))*180/pi;
}
```

We assume, for simply, that we get as results exactly four angles in the above algorithm for each i=2 to n. We have to run the above *Safe_Courses*(i) procedure for each pair: the own ship and the target ship number i (for i=2 to n) and as the result we get all safe sectors < $gamma_i1$, $gamma_i2$ >, <$gamma_i3$, $gamma_i4$>.

Next, we execute the *Common_Safe_Sectors* procedure for all target ships as the angle intersections of all safe sectors < $gamma_i1$, $gamma_i2$ >, <$gamma_i3$, $gamma_i4$> (for i=2 to n). The details of step one can be found in [12]. Let's denoted by $gamma_j$ elements of common safe sectors.

V. VERIFYING

Testing of NAVDEC on m/f Wolin was carried out on open sea in the period of four months.

There were following aims to verify during testing period:

1. Correctness of encounter parameters – to be verified by ARPA and Full mission simulator.
2. Correctness of new courses (which lead to pass other targets on presumed CPA) calculation – to be verified by radar and Full mission simulator.
3. Reaction of the system for changing initial settings – to be verified by Trial manoeuvre.

Next few figures present screenshots from NAVDEC interface.

During four months m/f Wolin made almost hundred voyages from Świnoujście to Tralleborg. During each voyage there were tens of collision situation (actual CPA were smaller than safe, presumed CPA). One of this is presented on figure 5.

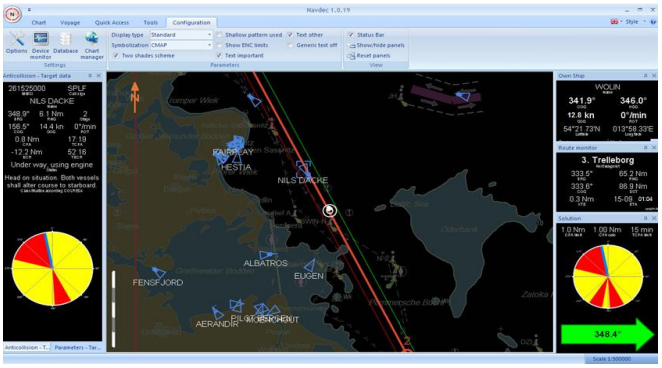


Fig.5. Encounter situation with m/f Nils Dacke.

In the situation presented on figure 5, CPA for target Nils Dacke is 0.8 Nm (Nautical mile). NAVDEC qualified encounter situation as "head on situation". According to Rule 14 of COLREGs, both vessels have to alter course to starboard to avoid collision.

Basing on safe, presumed by navigator CPA, system worked out and presented in the form of rosette sectors of safe courses (yellow) and sectors of dangerous courses (red). If own vessel take course from red sector and other target will keep her parameters, it will not be possible to pass on safe distance, and in critical situation collision vessels can collide. If own vessel take course from yellow sector, vessels will pass each other at least on presumed, safe distance.

The rosette shown in the left hand, down corner is an individual rosette for target, which was selected (by clicking on the chart) by navigator. The rosette presented in the right hand, down corner shows solutions for all targets within determined by navigator distance from own vessel (during tests in was 8 Nm). It's different that individual one, because two other targets are taken into account in calculation process. Present course of m/f WOLIN (341.9^0) is within red sector, this is why system NAVDEC displayed also suggested, optimal course. It's presented in the right hand, down corner in the form of green arrow with printed course (348.4^0), which enables both vessels to pass on presumed 1 Nm, which is in accordance of COLREGs and requires smallest deviation from presents course. Additionally, color of the arrow suggests to navigator direction of deviation i.e. green means "to starboard".

The figure 6 presents system "behavior" in dense traffic situation. During each voyage, m/f Wolin crosses twice Traffic Separation Scheme (TSS).

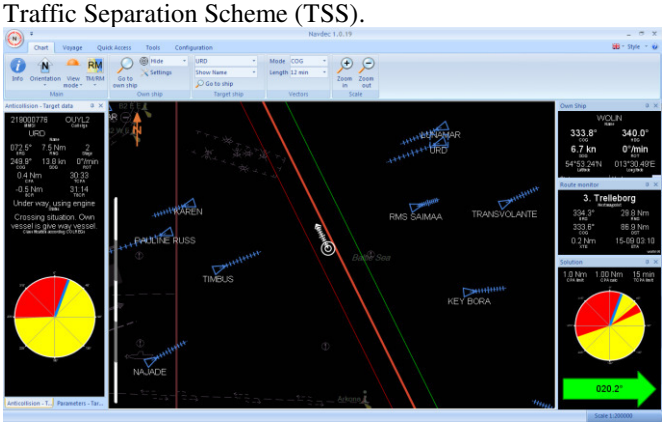


Fig.6. Dense traffic situation.

In this situation, CPA in relations to „Urd” (report displayed) and „Lunamar” is smaller than presumed 1 Nm. Additionally others targets should be taken into consideration when planning anti collision manoeuvre.

Individual rosette of target „Urd”, dangerous sector is arround 110^0 (from arround 270^0 to 20^0). On the common rosette, that sector is gradually increased both from left side (target „Lunamar”) or starboard side (targets „Timbus” and „RMS Saimaa”). According to the COLREGs, own vessel is give way vessel in relation to targets „Urd” and „Lunamar”. System NAVDEC has qualified encounter situation as „crossing situation” (Rule 15 of COLREGs) and suggested to alter course to starboard on course 020.2^0 (green arrow). Such manoeuvre enables to pass with all targets within 8 Nm, on the safe, presumed distance of 1 Nm.

VI. Results

Testing of NAVDEC on m/f Wolin was carried out on open sea from December 2014 to April 2015. Results, in general, are positive. In details system correctly calculates encounter parameters like CPA and TCPA. Displayed parameters were each time compared with ARPA. Additionally CPA and TCPA calculated by NAVDEC were compared with encounter parameters calculated by Full mission bridge simulator. Results show that NAVDEC is more precisely than ARPA particularly when ships are manoeuvring. In the first phase of manoeuvre CPA and TCPA presented by ARPA are useless and should not be taken into account in evaluation of encounter situation as it could lead of its misjudgment. Moreover NAVDEC informs navigator that targets have started their manoeuvres. In such situation target ship is flashing yellow. This function is not available in ARPA. Moreover:

1. In all verified cases own and target data, as well as encounter parameters presented by NAVDEC were correct,
2. In all verified cases qualification of encounter situation done by NAVDEC were according COLREGs,
3. In all verified cases NAVDEC correctly calculates and presents suggested, optimal anti collision manoeuvre,
4. In all verified cases NAVDEC correctly calculates and presents sectors of safe anti collision manoeuvres.

VII. Implementation

NAVDEC will find application on vessels and in land-based centres as an independent system or a module added to the existing navigational systems. Its main areas of use include:

- navigation-related decision support in collision situations – shipboard decision support system installed on the navigational bridge of merchant vessels (sea-going and inland shipping) and leisure boats (e.g. sailing ships, motor yachts)
- navigational decision support in collision situations – component of land-based vessel traffic services systems (VTS, VTMS, VTMIS, RIS)
- analysis and assessment of marine accidents at sea and on inland waterways – a system intended for experts working for maritime courts
- marine officer training centres offering courses in the Collision Regulations – a module of navigational simulators (e.g. ship-handling, ECDIS)

At present, the system is developed for operation in the open sea, so emphasis is put on developing functionalities for navigation in restricted waters. This is related to issues such as navigational restrictions of a water area and requires applications of advanced methods and tools of dynamic optimization.

It is also planned to develop versions of the system for marine training centres for ship-handling simulators and in longer term system functionalities for VTS centres will be extended with functionalities aiding vessel traffic control and management.

An example of analysis of marine accidents is presented below.

Based on the data included in the report [13], a simulation was made to determine parameters of the encounter and to generate possible anti-collision manoeuvres at certain moments of time. The solution does not account for manoeuvring components (kinematic equations). Figure 7 presents a reconstructed situation at 0900 hrs. The range of courses that assure safe passing at the preset CPA or larger is marked yellow on the circle. The recommended manoeuvre is indicated as ‘NEW COURSE’ and enables the ships to pass each other at the assumed CPA. The speed range satisfying the assumed criteria is marked green, and proceeding at ‘NEW SPEED’ will result in the ships’ distance during passing being equal the assumed CPA. At operator’s request, the system can display the recommended trajectory based on the generated solutions and the next waypoint (Figure 8).

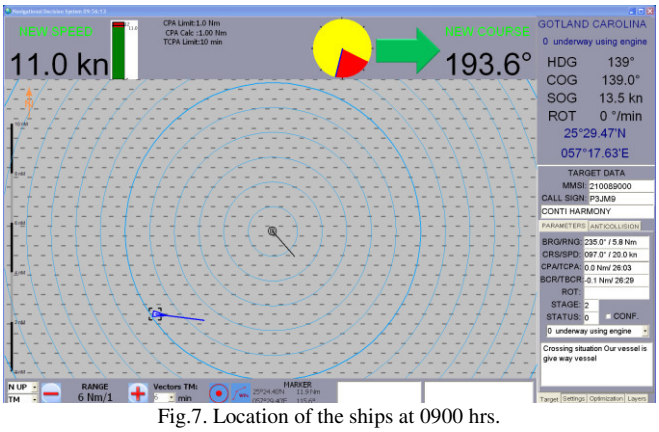


Fig.7. Location of the ships at 0900 hrs.

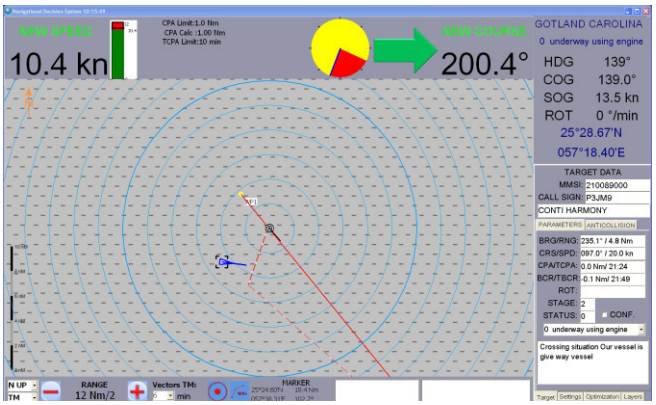


Fig.8. The recommended trajectory.

When the recommended manoeuvre is performed by own (system operator's) ship, the system assesses the situation as safe (green ship contour – Figure 9), as all the criteria have been satisfied. At the same time, in line with COLREGs, the situation remains qualified as before, so our (operator's) ship is still the give-way vessel.

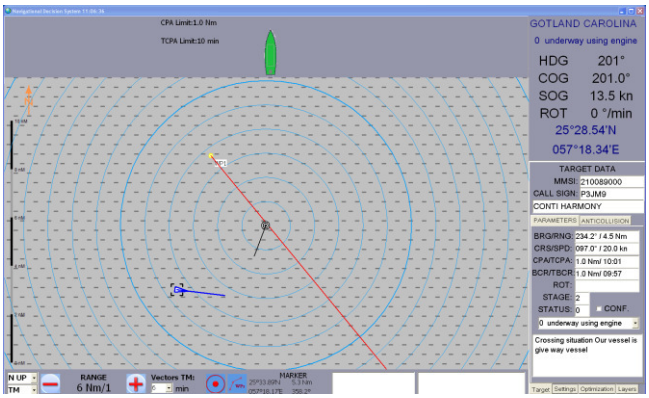


Fig.9. Situation after an anti-collision manoeuvre.

If the navigator does not take a preventive action, the system will continue to work out manoeuvres to be performed. If a collision cannot be avoided by altering course to starboard or by changing speed, proposed course alterations to port will be displayed (Figure 10).

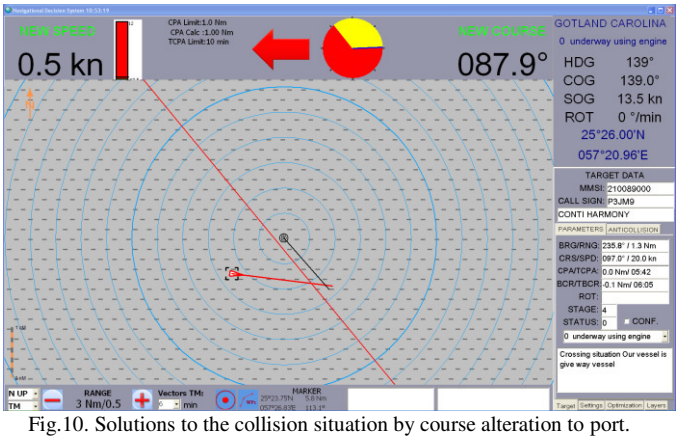


Fig.10. Solutions to the collision situation by course alteration to port.

At the time the ships come to a point where passing at distance of 1 Nm will not be possible, the system automatically reduces the assumed CPA by half. The new CPA taken into account while generating an anti-collision manoeuvre is displayed at the top screen denoted by CPA Calc (Figure 11).

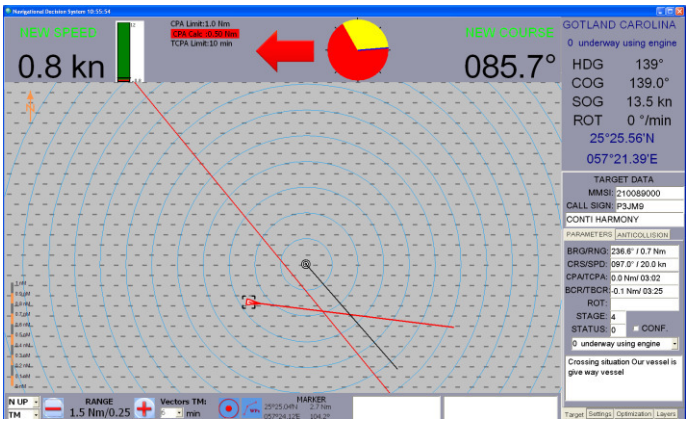


Fig.11. The manoeuvre generated after a reduction of the assumed CPA.

Future trend innovation leverage will be the coupling of NAVDEC with the ship's operational systems and equipment to enable automatic anti-collision manoeuvres. With such extension, the system will become a principal component of unmanned vessel control system. [14]

System is already installed on 13 ships.

VIII. Conclusion

According to the reports from the States in Baltic region there were 149 ship accidents in the HELCOM area in 2012 (Figure 12), which is 6 more than the year before (increase of 4%) and 19 more than in 2010 (increase of 15%).

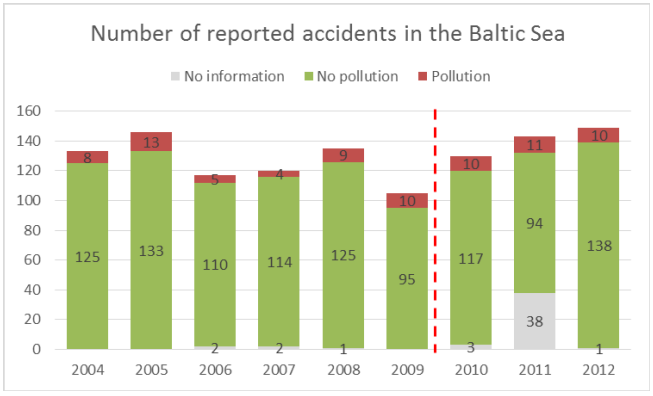


Fig. 12. Accidents in Baltic region in the period 2004-2012[4].

On the figure 13, there are statistical information from insurance company The Swedish Club.

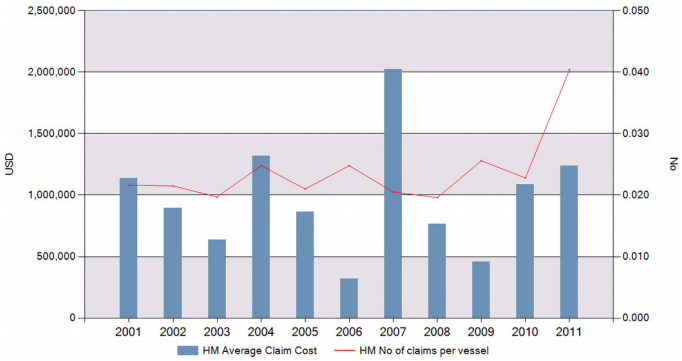


Fig. 13. Average claim cost & frequency 2001 — 2011, limit >= USD 10 000 [2]

According data presented above, average cost of collision is more than 1 million USD. The Swedish Club shares 13.6% (2010) of hull and machinery insurance global market. According Figure 5, around 2.5% of vessels are in collision every year i.e. over 1,900. In this situation the total cost of collisions is around 2 billion USD per year. According data from International Union of Marine Insurance, worldwide premium volume in 2013 was 34.2 billion USD. [14]

NAVDEC gradually increases functionality of existing navigational systems [15]. First of all it qualifies encounter situations according COLREG. Navigator receives notification if she is stand-on or give way vessel and from which rule it comes from. Moreover system gives ready solution of collision situation i.e. save courses and speeds which enable to pass with other targets on assumed CPA. Additionally suggested trajectory is presented on the chart [16].

If mandatory installation of a navigational decision support system reduces number of collision only by 1%, total savings, only for insurers, will be around 20,000,000 USD per year. The collision between Gotland Carolina and Conti Harmony is a proof that this percentage will be much higher.

IX. Future work

Author presented the results of verification of decision support system in real condition. The system involves a number of simplifications for the meeting stages, which have a place in the decision making process. In addition, implementation of the algorithm in the NAVDEC system will require taking into account the limitations of restricted area during the manoeuver and moving away from the assumption that ships can only manoeuver by course and not by course and speed. Besides, the key point of the proposed algorithm, selected by the frame will require the use of the heuristic method with the low time complexity. In the first step author will attempt to apply solutions based on ant colony algorithm described in [17], game control [18], genetic algorithms described in [19] and [20] or evolutionary algorithms described in [21] or ant colony algorithm described in. Exit condition of the loop will also require the use of fast solutions in the field of computational geometry.

References

[1] http://albumpolski.pl/artykul/polskie-statki-cz-2
[2] The Swedish Club, *Collisions and Groundings*, www.swedishclub.com
[3] Project Horizon 2012 Research Report, www.project-horizon.eu
[4] HELCOM Baltic Marine Environment Protection Commission, *Shipping accidents in the Baltic Sea in 2013* http://helcom.fi
[5] Sub-Committee On Navigation, Communications and Search and Rescue, e-Navigation Strategy Implementation Plan, January 2015
[6] J. Magaj, P. Wołejsza, Algorithm of working out anticollision manoeuvre by decision-supporting system, Advanced Computer Systems (ACS 2008), Międzyzdroje 2008.
[7] J. Magaj, P. Wołejsza, Analysis of possible avoidance of the collision between m/v Gotland Carolina and m/v Conti Harmony, Annual of Navigation No 16, pp 165-172, Gdynia 2010.
[8] Z. Pietrzykowski, P. Borkowski, P. Wołejsza, NAVDEC – navigational decision support system on a sea-going vessel, Scientific Journals Maritime University no 30 (102), pp. 102-108, Szczecin 2012.
[9] J. Koszelew, P. Wołejsza, Last minute manoeuvre as a part of maritime transport logistic system, Logistyka, No 4, 2014.
[10] A. Lisaj, The Method of the Navigation Data Fusion in Inland Navigation. Marine Navigation and Safety of Sea Transportation, Navigational Problems, pp.187-193, CRC Press 2013.
[11] A. Lenart, Manoeuvring to required approach parameters – CPA distance and time. Annual of Navigation 1/1999.
[12] P. Wołejsza, Multi-agent decision support system in collision situation, dissertation, Maritime University of Szczecin, 2008.
[13] Danish Maritime Administration, Casualty investigation reports, www.dma.dk
[14] Z. Pietrzykowski, P. Borkowski, P. Wołejsza, Maritime Intelligent Transport Systems, Communication in Computer and Information Sciences, Telematics in the Transport Environment, Springer Verlag Berlin Heidelberg, pp. 284-292, 2012.
[15] P. Wołejsza, Functionality of navigation decision supporting system - NAVDEC, Marine Navigation and Safety of Sea Transportation, Navigational Problems, pp. 43-46, CRC Press 2013.
[16] J. Koszelew, P. Wołejsza, Anticollision manoeuvre optimization in the NAVDEC system, 2014, Advanced Computer Systems (ACS 2014) Przegląd Elektrotechniczny No 2/2015 pp. 27-30 http://pe.org.pl/abstract_pl.php?nid=9004
[17] Ming-Cheng Tsou, Chao-Kuang Hsueh, The study of ship collision avoidance route planning by ant colony algorithm, Journal of Marine Science and Technology, Vol. 18, No. 5, pp. 746-756 (2010)
[18] J. Lisowski, The sensitivity of computer support game algorithms of a safe ship control. International Journal Applied Mathematics and Computer Science, Vol. 23, No 2, 2013, pp. 10.
[19] J. Koszelew, K. Ostrowski, A genetic algorithm with multiple mutation which solves orienteering problem in large networks, Computational Collective Intelligence - Berlin : Springer-Verlag, 2013, 356-365
[20] R. Szłapczyński, J. Szłapczyńska, On evolutionary computing in multi-ship trajectory planning, ApplIntell (2012) 37:155–174, DOI 10.1007/s10489-011-0319-7
[21] R. Śmierzchalski, Z. Michalewicz, Modelling of a ship trajectory in collision situations at sea by evolutionary algorithm. IEEE Transaction on Evolutionary Computation, Vol. 4, No. 3, 2000, 227–244.

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 53–60

DOI: 10.15439/2015F288
ACSIS, Vol. 6

Unsupervised Extraction of Graph–stream Structure for Purpose of Knowledge Retrieval and Information Fusion

Radosław Z. Ziembiński
Poznan University of Technology
Faculty of Computing
Piotrowo 3, 60–965 Poznan, Poland
Email: radoslaw.ziembinski@cs.put.poznan.pl

***Abstract*—Technologically inevitable introduction of various kinds of sensors to our life resulted in the production of huge amount of data delivered as streams. An improper acquisition of information may lead to errors caused by mixing observations coming from different processes threads. Some remedy can bring a proper representation of information. Hence, this paper introduces a graph–stream structure representing performance of complex multi–threaded process. The proposed network representation can separate information describing multiple threads and allows for modeling causal relationships between them. It gives separated and segregated information opening opportunity for development of qualitatively better and simpler knowledge retrieval algorithms. Further, the paper delivers a method for this representation extraction from multivariate data stream. It would be done by a clustering algorithm particularly designed for this purpose and evaluated quantitatively and qualitatively on example sets of data.**

I. Introduction

A RECENT decade revealed even more profoundly how modern life interleaves technology and habits. Miniaturization gained momentum when the processing performance and storage capacity (in relation to price) reached a threshold allowing for inexpensive execution of complex algorithms used in knowledge retrieval and machine learning. Then, it became possible to adjust functionality and interface of the device to the user expectations in semi–automatic manner using machine learning methods.

Modern smart devices have to handle big amounts of data and provide feedback to the user in a reasonable time. Services quality usually depends on the performance of the knowledge retrieval. Unfortunately, the deadline put on the processing time is rather difficult to met by many categories of data mining algorithms developed for stationary data sets. Particularly it happens, if the search space exponentially depends on the input size.

Another difficulty may arise if the observed process is a complex one (e.g., multi–threaded) but the sensor used to make observations is relatively simple. Then, the acquired data stream contains information which is a superposition of observations coming from different objects. Direct handling by common motifs finding or patterns mining algorithms may be difficult in the case. Simply speaking, we cannot identify particular objects and separate them easily for the mining purpose. Thus, even slight lack of synchronization in the recorded objects behavior can reorder parts of data. It may lead to low quality results whose application poses risks. An example ambiguity introduced by such distortions is illustrated on Fig. 1, where a small shift in events order changes historical circumstances. In the context of information fusion, it is more difficult to benefit from synergy in observations made by different kinds of tools. It happens, because direct comparison of context is not possible after the "collapse" of information collected from differently synchronized threads.

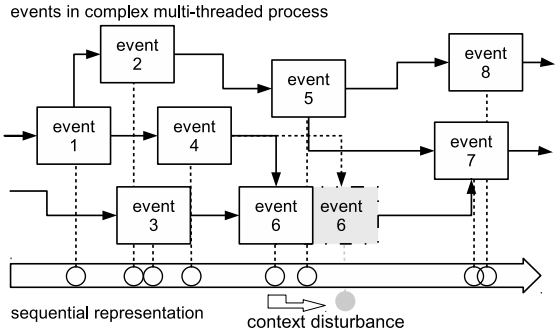


Fig. 1. The structural context "flattening" after the serialization of information.

The confidence in the processing can be improved if the simple sequential representation will be replaced by a vessel crafted particularly for the purpose of knowledge retrieval. Coming from such motivation, this paper provides study of the research on a new graph–stream data structure. An introduced container is a network graph capable of storing separated information about episodes observed in the process. Its nodes carry segregated information about events while edges can describe complex causal relationships between them. According to author expectations this structure should solve some from above issues. Mainly, these related to unreliable context construction where events mixture describing different threads try to define causal relationships.

Looking at literature, knowledge retrieval focuses partic-

ularly on the patterns mining in data streams. The context separation problem can be partially solved, if data are converted to chronologically ordered set of sequences. Then, each sequence may describe a single subprocess. Such method can be supported by various data mining algorithms [1], [2], [3], [4], [5]. Even, if this approach seems to be better from the single stream approach, this representation still loses information about conditional dependency between constituent sequences. A similar approach has been used also in the context of multivariate data sets [6].

The conditional dependency can be preserved, if sequential representation would be replaced by relevant multidimensional data structure. Particular interest raises graphs as natural extension of sequential representation on multidimensional ones. However, ways to construct graphs from sequential data can be very diverse. The mining of the graph–based representation has been studied for stationary and non–stationary graphs e.g., [7], [8], [9]. Unfortunately, exploration of general graphs is computationally expensive since verification of graph isomorphisms is non–polynomial [10] (but sub–isomorphisms is NP–complete [11]). These considerations lead to the conclusion that the compromise solution should be sought in specific families of graphs e.g., directed acyclic graphs. In this spirit, there have been proposed methods for mining partial orders in sequential data [12], [13].

This paper contributes to the state of art by an introduction of the graph–stream structure and algorithm for its extraction from observations collected by set of sensors. It has a following outline. A next section introduces the graph–stream definition. Then, the paper delivers the extraction algorithm which can be used to obtain the graph–stream from the stream of observations. A following section contains a presentation of results from the algorithm performance evaluation on artificial data sets. Reported experiments have involved finding frequent episodes in the graph–stream and measurement of their autocorrelation. Remaining part focuses on conclusions.

II. Graph–stream definition

The proposed data structure is intended to store information about observations of a complex process. It is assumed that the observed process includes few subprocesses that produce stimulus to sensors simultaneously. Alternatively, the process is observed by different sorts of sensors from various perspectives at once. In a consequence, the acquired data stream usually contains a mixture of information generated by observed subprocesses. Introduced structure and algorithm should keep it separated for the purpose of the following processing.

It was already mentioned that the graph–stream is the customized directed acyclic graph. Let's begin its introduction from a definition of some basic carriers of information. In our case, the processed data stream is described by a set of nominal or continuous attributes A.

Subsets of A determine informational content of graph nodes. Node $n_{p_{type},i}(p_{ts}, A_i) \in N$ is a unit of information collected at one moment. It is described by type p_{type}, event occurrence time–stamp p_{ts} and attributes' values $A_i \in A$. Directed edge $e_{i,j}(n_i, n_j) \in E$ connects nodes n_i and n_j and represents causality relation between nodes. It uses the symbol $\rightarrow$.

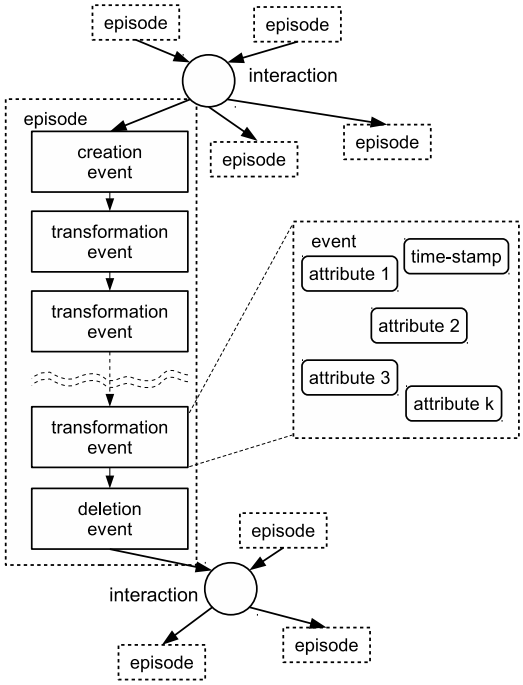


Fig. 2. The graph–stream data representation.

Above definitions lead to formulation of the directed acyclic graph. A directed cycle is a sequence $n_1 \rightarrow n_2, ..., n_{k-1} \rightarrow n_k \in N$ of nodes collected along a path made from directed edges where $n_1 \equiv n_k$. $G_{DAG}(N, E)$ is a directed graph where the cycle is absent.

The graph–stream uses distinct types of nodes to describe static and dynamic properties of observed objects:

- State node $n_{S,i}(p_{ts}, A_i) \in N$ collects all values of attributes describing state of a particular observed object in the process. This node is used to represent a boundary state of the episode just after its creation or deletion. However, it can be produced from the interaction, too.
- Transformation node $n_{T,i}(p_{ts}, A_i) \in N$ describes a modification of a single observed object. However, it is self modification without external influences. This node represents the transformation event.
- Interaction node $n_{E,i}(p_{ts}, A_i) \in N$ describes an observed incident involving an interaction between objects from different threads. At this moment episodes that went into the interaction event collapses and produce a new non–empty set of episodes. It emerges from this definition, that node binds causally set of interacting episodes to their products. This node has only the time–stamp attribute.

There is no constraint in this proposal shaping attributes distribution between state and transformation nodes. It is only a suggestion to keep boundary states information in

state nodes and operations or changes of the episode state in transformation nodes.

Above definitions have to be complemented about constraints imposed on edges. The interaction node joins episodes by causal relation. According to graph–stream's structural assumptions it can connect only a non–empty set of deletion nodes to a non–empty set of creation nodes. It is many–to–many relation reflecting causality dependencies between episodes. Relationship between transformation and state nodes have simpler interpretation. They can be connected only by single ingoing and outgoing edges. Hence, they form a sequence beginning from the creation node, lasting through transformation nodes and ending with the deletion node. It is called the episode (body).

Finally, let us define *the graph–stream* $G_{ST}(N, E)$ as a set of episodes S connected by interaction nodes according to above constrains imposed on edges. Illustrations of introduced structures are drawn on Fig. 2 at different levels of details.

III. Graph–stream Extraction Method

A. Episodes identification

Algorithm proposed for the graph–stream extraction converts data stream in three subsequent phases. Due to complexity and size of the algorithm code it is difficult to describe all details. Hence, this description primarily put attention on the most important design features. It should give sufficient hints about algorithm structure to understand its construction. For matter of convenience, the algorithm will be called *GATAC* (GrAph–sTreAm extraCtor).

The first phase of the processing is handled by a dedicated on–line clustering algorithm. Its overall pipeline for processing new data objects and performing the graph–stream extraction presents Fig. 3.

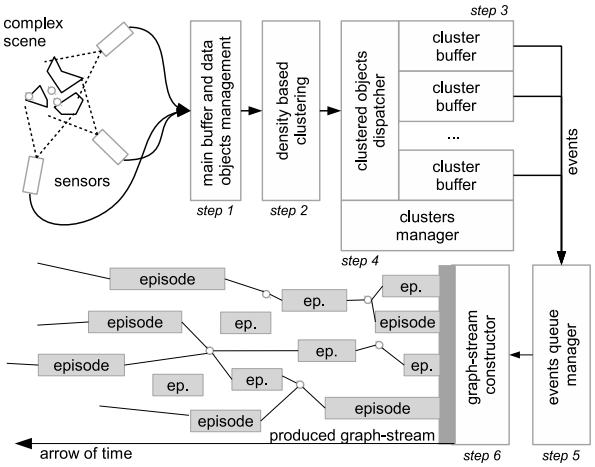


Fig. 3. The process of the graph–stream extraction.

Proposed implementation of the algorithm uses a fixed set of attributes A. However, it does not enforce their full usage for each processed data object. This clustering is performed within groups of data objects described by the same sets of attributes (subsets of A). In a consequence, this method can construct the graph–stream from separate sources of data even if they share some or neither attributes.

Let's now describe life–cycle of a single data object processed by this algorithm. At the beginning, values of data object attributes are normalized. It is a necessary initial step because later a similarity function aggregates partial similarities calculated for compared attributes. The normalization is done according to intervals approximating attributes domains ranges (updated on–line). Then, incoming data objects are sorted in B according to values. If the main buffer appears to be full then the oldest data object is removed from B to make a free space for the new one (step 1, Fig. 3).

After placing the data object in B, the algorithm begins sending messages to objects in the neighborhood to determine the set of the most similar neighbors. Messages that are circulating between data objects are stored in two structures: a message queue and replies sorted list. Those from new data objects are stored in message queues of neighbors. In this implementation all structures storing messages have fixed sizes to delimit the memory usage. Limits prevent from accepting too many messages from other data objects at the cost of neighborhood identification accuracy (step 2).

The procedure is performed in several subsequent iterations to determine the neighborhood of specified size. By this way the algorithm performs iteratively the neighborhood search in the breadth–first manner.

$$
\begin{aligned}
\sigma_o(o_1,o_2) &= (\delta_d(o_1,o_2) * \delta_t(o_1,o_2) \\
&\quad *\delta_l(o_1,o_2))^{1/3} - 1 \\
\text{where:} & \\
\delta_d(o_1,o_2) &= 1 + (\textstyle\sum_{i=1..|A_{o_1,o_2}|} |o_1.A[i] \\
&\quad -o_2.A[i]|)/|A_{o_1,o_2}| \\
\delta_t(o_1,o_2) &= 1 + (|A_{o_1,o_2}| * |o_1.tstamp \\
&\quad -o_2.tstamp|)/(|B| * |B.tspan|) \\
\delta_l(o_1,o_2) &= 2 - (o_1.labels \cap o_2.labels) \\
&\quad /(o_1.labels \cup o_2.labels)
\end{aligned}
\tag{1}
$$

Afterward, the algorithm browses all non–empty message queues and prepares replies to their senders. They contain information about similarities between pairs of data objects. The range of replies is delimited by a decreasing skip counter. In a consequence, only the closest neighbors are visited by replies sent to new data objects.

Equation 1 is used to calculate similarity between two data objects. If it is equal to 0, then objects are identical. The similarity computation takes into account three properties of data objects pair i.e., difference of attributes values $\delta_d(o_1, o_2)$, timestamps $\delta_t(o_1, o_2)$ and labels $\delta_l(o_1, o_2)$. Variables and constants used in the equation have following explanation: set of common attributes for pair of objects $A_{o_1,o_2} \subseteq A$, data object's time stamp $o.tstamp$, main buffer B, time span $B.tspan$ (calculated on–line from timestamps of stored data object's), labels associated to data object $o.labels$. Labels are nominal values that can be used for further differentiation of data objects by describing e.g., data source properties.

After the sender data object (the new one) received all replies from close neighbors, a swapping procedure begins to filter out some of them. This operation is performed to preserve smoothness of the cluster distribution. In the result, data objects from subspaces containing different densities become better separated, even if they adjoin. It prevents from merging sparse clusters to denser ones (if the density proportion is above the parametrized threshold) and excludes noise.

If the replies list becomes empty, then the new data object receives a new cluster identifier. However, the new cluster buffer is created only if the second object appears with the same identifier. Hence, a standalone object does not invoke creation of the cluster buffer and new thread. It makes the processing more efficient by eliminating noise.

At this moment, data objects are moved to the cluster buffer assigned to a single thread that can produce one or more subsequent episodes (step 3). An event signaling the cluster creation is thrown if its population passes the parametrized threshold. The threshold ensures that the statistics from data distribution in the cluster buffer are robust.

The newly added data object can merge two or more clusters if both are located in its neighborhood (step 4). The algorithm maintains alteration counters assigned to cluster buffers. If their values become greater than threshold Θ_C, then the algorithm begins a breadth–first introspection of the main buffer. It is performed according to neighborhood information stored in replies lists of objects. This procedure rewrites all identifiers of data objects from clusters that become connected since the previous introspection. The new cluster identifier value is taken from the most populous contributor to preserve the strongest supported thread. This operation throws interaction event, if the merger occurs.

If the main buffer is full, then the algorithm removes the data object according to FIFO rule. Removal procedure also modifies the alteration counter associated to the cluster. So, it may trigger the breadth–first introspection of the main buffer, too. If it happens, a fragmentation of the cluster may be revealed. Then, each new fragment of the original cluster receives distinct new cluster identifier, while the most populous one retains the previous one. Additionally, a relevant interaction event is produced.

The cluster buffer describing episode may be discarded if its support terminates. It happens, after it has not been supported for a time longer than average "time distance" between data objects already stored in the cluster multiplied by the parametrized threshold. Such termination procedure facilitates clusters supported at different rates by input data. Moreover, it is robust to a slow drift of the support frequency.

B. Detection of transformation events

The detection of transformation events is necessary to construct the episode body. It uses the cluster tracing mechanism to detect significant changes reflecting shifts in data distribution. At this stage, it can be done relatively simply due to the fact that the clustering phase binds each episode to the life–cycle of a single cluster.

The data drift tracing begins just after sending of the interaction event related to the episode creation. After this event, the algorithm can calculate plausible statistics and send following transformation events notifying about changes in the data distribution. To prevent from unnecessary recalculation of statistics, the algorithm uses modification counters assigned to clusters. The counter is incremented each time when new data object is added or removed from the cluster buffer. Statistics become recalculated if it passes Θ_C.

Algorithm 1 Procedure for transformation events detection.

Require: Set of clusters - C, set of modified statistics - M, mean threshold - Θ_{mean}, standard deviation threshold - Θ_{mean}, density threshold - Θ_{dens}
Ensure: Generated data object - o

```
for all c ∈ C do
  if c.modificationsCounter > Θ_C then
    c.modificationsCounter = 0
    M = c.calculateStatistics(Θ_ang, Θ_mean,
    Θ_std, Θ_dens)
    if M <> ∅ then
      c.updatePreviousStatistics(M)
      c.fireTransformationEvent()
    end if
  end if
end for
```

In the current implementation, calculated statistics include means and standard deviations calculated for each attribute alone. Additionally, the event may include average density of the cluster. Their calculation is performed according to Alg. 1. Current calculated values are compared to previously determined state ps and the reference state rs reported in the preceding event. If statistics pass adequate thresholds (mean, deviation and density) for attributes, then the procedure sends the transformation event with the current state (or optionally information about changes). The step also involves replacement of modified statistics from the reference state by current ones. Hence, supporting procedure detects drift occurring for linear changes in data statistics. It reduces output information by preventing from a flood of following events delivering information about linear shifts in data distributions.

The construction of the graph–stream is a simple step. Events contain full information about predecessors, successors and episodes affiliation inherited from cluster buffers. It is used for building a graph–stream just by extending a set of episodes tails by incoming events (step 5 and 6). At interaction, participating episodes are closed and bound to result in a form of newly created tails of following episodes. For this purpose, the proposed algorithm uses a map of episodes M that allows for fast manipulation of them.

C. Computational complexity and scalability of the extraction algorithm

The computational cost of the algorithm depends on data dimensionality, sizes of the main buffer and cluster buffers.

TABLE I
PARAMETERS OF THE GRAPH EXTRACTION ALGORITHM.

Parameter:	$\|B\|$	$\|C\|$	Θ_ρ	Θ_σ	Θ_C	Θ_T
Value:	500	40	2	2	16	2
Parameter:	Θ_{mean}	Θ_{std}	Θ_{dens}			
Value:	0.1	1.5	1.5			

The data in the main buffer are sorted thus the complexity is bound to $O(|A| * log(|B|))$, where $|B|$ is the size of the main buffer. Its size determines the assignment of new data objects to cluster buffers and data distribution "forgetting". On the other hand, size of each cluster buffer is relatively small. It has to be sufficiently large for accountable statistics calculation (few dozens of data objects). Data objects update procedure requires $O(|A| * |B|)$ steps for maintaining messages and replies. However, the calculation of statistics requires a double loop on each cluster's data set. Therefore, the complexity is $O(|A| * |C|^2)$ where $|C|$ is the size of the cluster buffer. Fortunately, statistics are recalculated only if the cluster's modification counter becomes greater than the threshold Θ_C. Hence, the rate of statistic recalculation is delimited by $1/\Theta_C$. The construction cost of the graph–stream is in order of $O(|A| * |C| * log_2(|M|))$ if all cluster buffers would send messages to neighbors.

Memory costs of the algorithm have been delimited by fixed buffers sizes. It refers to the main buffer $|B|$, clusters (buffers) $|C|$, both messages queues attached to data objects and used for the neighbors finding. The described implementation prefers the control on memory usage over the results quality. This design assumption allows its deployment on mobile or embedded devices. Of course, it is feasible until small buffer sizes would immerse the processing quality below acceptable level.

IV. EXPERIMENTAL EVALUATION

The purpose of experiments was to evaluate the process description stored in the extracted graph–stream. Described experiments were conducted on data obtained from real and artificially generated streams. Different kinds of experiments were performed to measure the algorithm's performance efficiency and its extraction accuracy. During them, a quality has been performed by measurement of episodes autocorrelation. The process of autocorrelations identification can be related to the problem of finding frequent item sets in nominal data stream [14], [15].

The extraction algorithm was evaluated in experiments with default settings on Table I. The main buffer size was set to $|B| = 500$ and clusters buffers sizes were delimited to $|C| = 40$.

A. Description of data sets

Artificial data sets are generated by procedure that produces data objects describing a group of interacting clusters. There are three modes of the generator described in Alg. 2. The data generation algorithm selects the cluster that produces data objects according to round–robin. Then, the generation is performed by distorting the cluster's center with a random value from Gaussian distribution.

Algorithm 2 Synthetic data generation algorithm.

Require: Clusters radixes - $crad$, time step - $cstep$ and variability - $tvar$, data object dimensionality - $|A|$, algorithm mode - $mode$, number of clusters - cno
Ensure: Generated data boject - a
{make a timeshift}
$iteration = iteration + 1$
$timeStep = randomUniform(cstep, cstep + tvar)$
$phase = phase + speed * timeStep$; $sPhase = sin(phase)$
$oidx = iteration \% cno$
if mode == 3 **then**
  $cobj = (int)(phase/PI)$
  if random(0, 1) ¡ abs(sPhase) **then**
    $oidx = cobj \% cno$
  end if
end if
{select cluster identifier}
$eidx = oidx/2 + 1$
$dirId = (oidx\%2 == 0)?(eidx) : (\ eidx)$
{calculate attributes' values}
for all d ∈ 1...$|A|$ **do**
  $dir = (dirId\&(1 << d))?(1) : (-1))$
  $dir = dir * (1.0 - crad)$
  if mode == 1 **then**
    $disp = randomNormalDist(0, crad)$
    $a[d] = (dir * sPhase + disp)/2 + 0.5$
  else if mode == 2 **then**
    $disp = randomNormalDist(0, 3 * (crad + 0.0001) * absT(sPhase))$
    $a[d] = (dir * crad * 2 + disp)/2 + 0.5$
  else if mode == 3 **then**
    $disp = randomNormalDist(0, crad)$
    $a[d] = (dir * crad * 5 + disp)/2 + 0.5$
  end if
end for
return a

The described generator has three modes of operation. For $mode = 1$, the generator forms hyper–spherical clusters with movement driven by sinusoid. They are moving forth and back, from edges to the center of hyper–box. If $mode = 2$ then it generates pulsating clusters. This process produces data objects intermixed in the middle of the hypercube in repeatable intervals. The final one is generated for $mode = 3$. It generates "blinking" clusters that gain and lose support periodically. In this case, clusters do not change their position or size but their density oscillates. These generated data streams show Fig. 4.

These distributions represent different kinds of interactions between observed objects. In the first case, the behavior of clusters leads to their complete coverage in the middle of the hypercube. Then, subprocesses become indistinguishable. The second stream describes subprocesses that interfere partially.

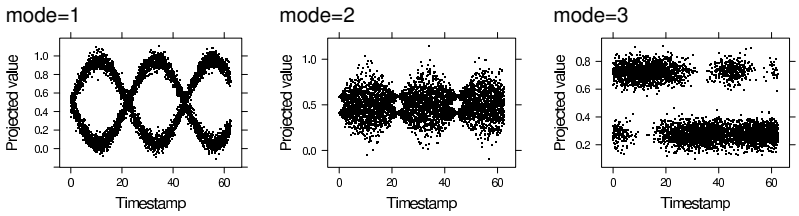


Fig. 4. Artificial data streams obtained from the generator.

The interference manifests itself as overlapped distributions. In the last case, there are no interactions between clusters at all. Observed subprocesses do not interact and their oscillations are relatively fast.

B. The processing performance measurement

Measurement of the processing efficiency has been done on artificial data stream containing 100000 data objects. It was generated for $mode = 1$. This experiment was performed on data streams generated for different clusters numbers at different dimensionality and processed with various main buffer sizes.

The conducted experiment included four sub–experiments from whom results are presented on Fig. 5. The top–left plot of Fig. 5 contains performance measurement of the graph–stream extraction algorithm for data of different dimensionality. This experiment was conducted for $a1 : |A| = 12$, $a2 : |A| = 9$, $a3 : |A| = 6$, $a4 : |A| = 3$ and $|B| = 500$. It can be noticed, that the number of dimensions have impact on the algorithm's iteration performance. It acknowledges the theoretical analysis of the main buffer and messaging performance. The main buffer has size $|B|$ and consists of $|A|$ sorted lists. A low cost of logarithmic access to B is reflected in results. The top–right plot delivers measurement of the computational costs for different main buffer sizes. Results were obtained for data dimensionality $|A| = 9$ and $b1 : |B| = 500$, $b2 : |B| = 800$, $b3 : |B| = 1100$ and $b4 : |B| = 1400$. Impact of main buffer size on the iteration cost is small because the binary search is used to get access to sorted objects. For both above experiments cno was equal to 2. All measurements include time required to store the produced graph–stream in memory. It explains slightly increasing costs during the processing.

Bottom plots describe the dependence of the extraction algorithm performance on data complexity. These experiments have been performed for stream carrying different number of clusters in the stream $c1, d1 : cno = 2$, $c2, d2 : cno = 4$, $c3, d3 : cno = 6$ and $c4, d4 : cno = 8$. The results were measured for $|A| = 9$ and $|B| = 500$. The bottom–left plot contains the processing cost measured per iteration, while the bottom-right one delivers cumulative numbers of generated events by the algorithm for the graph–stream extraction. Intensity of all data streams was the same. Therefore, the frequency of data objects per cluster was lower for streams delivering more clusters. It can be observed that the costs of the processing is the greatest for the smallest number of clusters at $cno = 2$. This result seems to be counterintuitive but it is caused by more intense messages forwarding. For a lower number of denser clusters the count of connections between data objects is relatively higher. Thus, the cost of handling communication is higher, too. Results form the bottom–right plot can be interpret according to intuition since it is clear that more clusters requires proportionally more events generated to describe them.

All experiments involving measurement of the performance were conducted on x64 workstation at clock 3 GHz. Each measurement was an average from 1000 following iterations. Microseconds scale oscillations observed on plots are presumably caused by hardware, operating system or C++11 standard libraries. All experiments have been performed on multi–core hardware. The affinity of processes to cores changed during the processing what might promote oscillations.

C. Similarity of episodes

Evaluation of autocorrelation requires a similarity measure that would allow for mutual comparison of episodes. Similarity computed for a pair of nodes takes into account differences in attributes values, the relative occurrence times (in the respect to timestamps of the episodes heading nodes) and densities. It is a geometrical mean from above partial similarities. To calculate partial similarity, attributes values difference dif is transformed with function $exp(-abs(dif) * weight)$. It standardizes results by scaling them using weights chosen per attribute. Then results are averaged and partial similarity of attributes is computed. The densities and occurrences times are threated separately as qualitatively distinct entities. But their differences are also standardized by that weighted exponential function. Weights help in the similarity function tuning. They can be used when we need to underscore particular property. Calculated similarity takes continuous values from 0 to 1. The value equal to 1 means that nodes are identical and related events have occurred in the same relative time measuring from respective episodes heads.

Episodes are evaluated according to fairly complicated procedure. Initially, pairs of state nodes beginning and terminating episodes are compared. Then, all possible pairs of transformation nodes from both episodes are enumerated to calculate their similarities. Afterwards, pairs are sorted according to similarity value and there are chosen ones with the greatest similarities. During this selection, it is forbidden to take two pairs containing at least one transformation node the same. In the result, only the strongest set of ties between the pair of distinct episodes survives selection. The selection is done

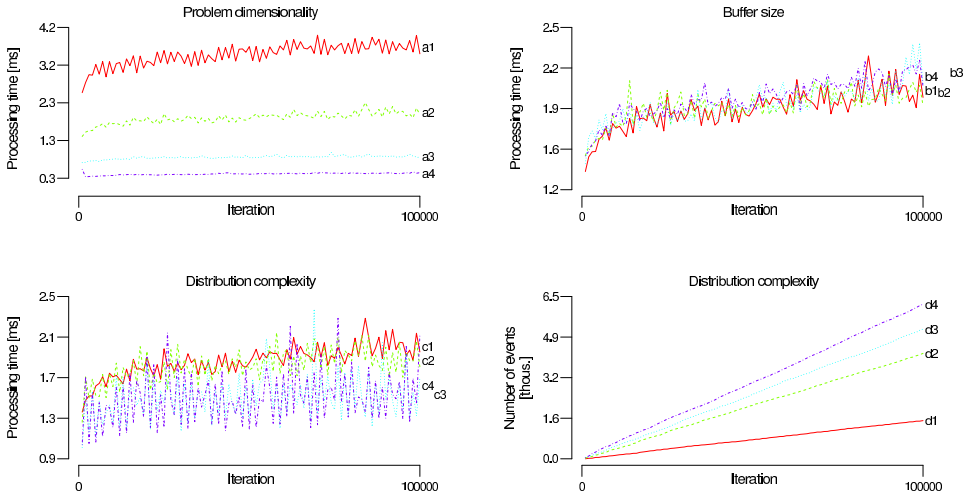


Fig. 5. The graph–stream extraction process efficiency.

regardless sequential order of transformation nodes in the episode. Fortunately, the order of nodes is taken into account elsewhere. Let's remind that the time–stamp difference is considered in events similarity calculation. Therefore, a pair of events occurred at different times in reference to their episodes heads time–stamps have low similarity and chance to be selected.

Global similarity is calculated as an average of similarities obtained from the comparison of pairs state and transformation nodes. It delivers a normalized result, where an episode compared to itself would receive the score equal to 1. The aggregation raises status of state nodes in relation to inner transformation nodes underscoring importance of border states of episodes. The introduced similarity function favors episodes containing the same attributes values in nodes and the same order of transformation nodes. Therefore, pairs of episodes that differ in duration would have lower similarity.

D. Autocorrelation of episodes in experiments

Results presentation begins from ones obtained for artificial data streams. These streams provided data describing 4 dynamic clusters (subprocesses) in 9 dimensions. Heat maps presents mutual similarities for episodes that belong to graph–streams and extracted from artificial data sets are presented on Fig. 6. They show only 200 of the most mutually similar episodes. Two figures for each graph–stream represent similarities of episodes (left) and contexts preceding them (right). Bottom legend contains information about correlation coefficient value between episodes and their contexts. Episodes are sorted according to their length (a number of transformation nodes) and the longest ones are located in the top–right corner of each heat map.

Fig. 6 reveals existence of similarities between episodes for periodic processes. There are observable groups of mutually similar episodes that have comparable sizes. They form "squares" located on the diagonal. They describe different stages of the subprocess evolution in a cycle and reveal characteristic lattice patterns inside. It is caused by the fact that data carry the description of 4 clusters that behave symmetrically. Thus, episodes describing them have very similar sizes and find their places in the same "square". Autocorrelation between episodes tells about a periodicity in the observed process and acknowledges the generator properties.

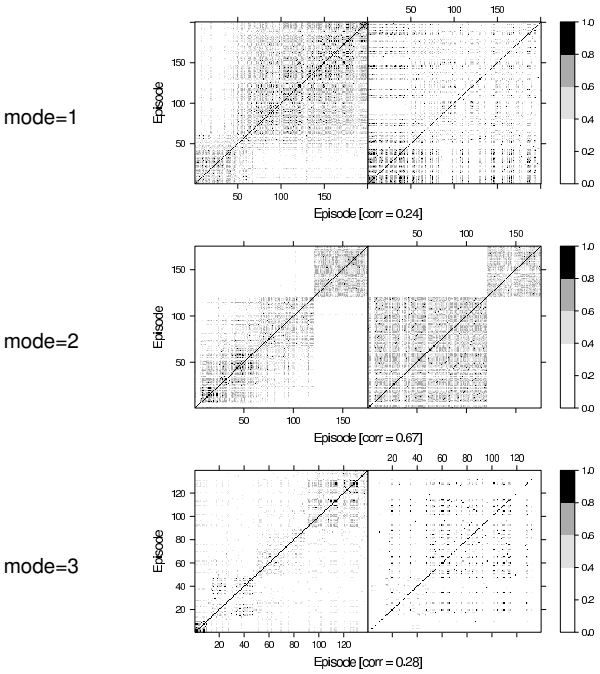


Fig. 6. Heat maps representing autocorrelation of episodes (the left column) and their the best contexts contributors (the right column) for artificial data streams.

Contributing contexts similarities are weakly correlated to related episodes. Aside from the artificial set generated at $mode = 2$, almost all correlation coefficients are very low. This suggests that the past context information has to be analyzed by looking deeper to the past. It is in line to

conclusions made for the sequential patterns mining where frequent elements of the pattern may be separated by many infrequent elements [1]. However, for some data sets e.g., ones generated at $mode = 2$ the correlation occurs using this algorithm even for adjacent episodes. Improvement in the cluster buffer tracing algorithm and casting transformation event may increase the efficiency of the correlated episodes finding.

Obtained results acknowledge that the exploration methods for finding patterns in graph–stream threads have to look deeper into past contexts of episodes. The "lattice–square" patterns on heat maps prove that the method correctly identifies episodes from distinct subprocesses. Because all experiments used the same parameter values therefore we can expect a further results improvement for algorithm parameters tuned to properties of particular data stream.

V. Conclusions

The proposed graph–stream extraction algorithm can produce directed acyclic graphs describing multivariate and multi–modal data streams. It has some advantages over commonly used sequential representation. In the first place, it can separate individual subprocesses within the observed complex process. This feature may be useful when we observe the process through array of heterogeneous detectors with one sink of data. Concluding, the data structure proposed in this paper may contribute in following areas to the state of art:

- It can help to identify undisturbed context of events. Events from different threads (subprocesses) become segregated and organized in episodes. Such structure is more robust when it comes to issues with synchronization for concurrent subprocesses. This allows for building simpler algorithms for knowledge retrieval and information fusion. This representation does not transform information stored in attributes. It just makes observations more understandable for the following processing by better organization.
- Network structure can be used to model complex causal relationships between events. Causality can be better reflected since DAG can represent relations many–to–many between dependent episodes. Preservation of attributes values from the original data stream and exposition of data dynamism would be useful in observation and analysis of dynamic processes.
- There are different sorts of nodes for representing interactions, transformations and border states. They form a grammar of a simple language e.g., like Feynman diagrams in physics this language is sufficiently capable to express a description of almost all discrete processes.
- It can lead to novel algorithms better exploring information about relationships between subprocesses in the complex process. This makes new opportunities for episodes clustering, classification, forecasting and correlations mining. In my opinion, a domain related to correlations mining between subprocesses is particularly interesting (regarding a subsequent information fusion). It may lead to new methods of information processing benefiting from synergy of data retrieved simultaneously from many qualitatively different sensors.

Acknowledgment

This paper is a result of the project financed by National Science Centre in Poland grant no. DEC-2011/03/D/ST6/01621.

References

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. of the Eleventh International Conference on Data Engineering*, ser. ICDE '95. Washington, DC, USA: IEEE Computer Society, 1995. ISBN 0-8186-6910-1 pp. 3–14.

[2] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *ICDE '01: Proceedings of the 17th International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2001, p. 215. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2001.914830

[3] H. Pinto, J. Han, J. Pei, K. Wang, Q. Chen, and U. Dayal, "Multi-dimensional sequential pattern mining," in *Proc. of the Tenth International Conference on Information and Knowledge Management*, ser. CIKM '01. New York, NY, USA: ACM, 2001, pp. 81–88. [Online]. Available: http://dx.doi.org/10.1145/502585.502600

[4] R. Ziembiński, "Algorithms for context based sequential pattern mining," *Fundam. Inf.*, vol. 76, no. 4, pp. 495–510, Dec. 2007.

[5] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire, and Y. W. Choong, "Mining multidimensional and multilevel sequential patterns," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 1, pp. 1–37, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1145/1644873.1644877

[6] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Causal information approach to partial conditioning in multivariate data sets," *Comput Math Methods Med.*, p. 17, 2012. [Online]. Available: http://dx.doi.org/10.1155/2012/303601

[7] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *Proc. of the 2002 IEEE International Conference on Data Mining (ICDM02)*. Washington, DC, USA: IEEE Computer Society, 2002, p. 721. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2002.1184038

[8] C. C. Aggarwal, Y. Li, P. S. Yu, and R. Jin, "On dense pattern mining in graph streams," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 975–984, Sep. 2010. [Online]. Available: http://dx.doi.org/10.14778/1920841.1920964

[9] A. Bifet, G. Holmes, B. Pfahringer, and R. Gavaldà, "Mining frequent closed graphs on evolving data streams," in *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 591–599. [Online]. Available: http://dx.doi.org/10.1145/2020408.2020501

[10] U. Schöning, "Graph isomorphism is in the low hierarchy," *J. Comput. Syst. Sci.*, vol. 37, no. 3, pp. 312–323, Dec. 1988. [Online]. Available: http://dx.doi.org/10.1016/0022-0000(88)90010-4

[11] J. R. Ullmann, "An algorithm for subgraph isomorphism," *J. ACM*, vol. 23, no. 1, pp. 31–42, Jan. 1976. [Online]. Available: http://dx.doi.org/10.1145/321921.321925

[12] H. Mannila and C. Meek, "Global partial orders from sequential data," in *Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '00. New York, NY, USA: ACM, 2000, pp. 161–168. [Online]. Available: http://dx.doi.org/10.1145/347090.347122

[13] R. Gwadera, G. Antonini, and A. Labbi, "Mining actionable partial orders in collections of sequences," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Springer Berlin Heidelberg, 2011, vol. 6911, pp. 613–628. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-23780-5_49

[14] J. H. Chang and W. S. Lee, "Finding recent frequent itemsets adaptively over online data streams," in *Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 487–492. [Online]. Available: http://dx.doi.org/10.1145/956750.956807

[15] M. Deypir and M. H. Sadreddini, "A dynamic layout of sliding window for frequent itemset mining over data streams," *J. Syst. Softw.*, vol. 85, no. 3, pp. 746–759, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2011.09.055

5th International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2015 - provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogenious field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

Topics

The topics of interest include, but are not limited to:

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information
- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

Event Chairs

Lasek, Piotr, University of Rzeszow, Poland

Paja, Wiesław, University of Rzeszów, Poland

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

Program Committee

Basarici, Samsun M., (Medical) Image Processing, Yasar University, Turkey

Deserno, Thomas M., Uniklinik RWTH Aachen University, Germany

Drahansky, Martin, Brno University of Technology, Czech Republic

Hashimoto, Hiroshi, Advanced Institute of Industrial Technology, Japan

Hassanien, Aboul Ella, Cairo University, Egypt

Iantovics, Barna, Petru Maior University, Romania

Kountchev, Roumen, Technical Univerity of Sofia, Bulgaria

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Kumar, Sajeesh, University of Tennessee, Health Science Center, United States

Majernik, Jaroslav, Pavol Jozef Safarik University in Kosice, Slovakia

Min, Fan, Zhangzhou Normal University, China

Olszewska, Joanna Isabelle, University of Gloucestershire, United Kingdom

Sawada, Hideyuki, Kagawa University, Japan

Shieh, Jiann-Shing, Dept. of Mechanical Engineering, Yuan Ze University, Taiwan

Sirakoulis, Georgios, Department of Electrical & Computer Engineering, Democritus University of Thrace, Greece

Strzelecki, Michal, Lodz University of Technology, Poland

Wtorek, Jerzy, Gdańsk University of Technology, Poland

Wysocki, Marian, Rzeszow University of Technology, Poland

Yanushkevich, Svetlana, University of Calgary, Canada

Zaitseva, Elena, University of Zilina, Slovakia

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 63–67
DOI: 10.15439/2015F393
ACSIS, Vol. 6

Consistency-Based Preprocessing for Classification of Data Coming from Evaluation Sheets of Subjects with ASDs

Krzysztof Pancerz, Aneta Derkacz
University of Management and Administration
Zamość, Poland
Email: kpancerz@wszia.edu.pl

Jerzy Gomuła
Cardinal Stefan Wyszyński University
Warsaw, Poland
Email: jerzy.gomula@wp.pl

***Abstract*—In general, the aim of our research is to adapt computational intelligence methods for computer-aided decision support in diagnosis and therapy of persons with Autism Spectrum Disorders (ASDs). In the paper, we are focusing on the data preprocessing step for cleaning a training data set for classifiers. An approach based on consistency factors is proposed.**

I. Introduction

AUTISM is a brain development disorder that impairs social interaction and communication, and causes restricted and repetitive behaviors, all starting before a child is three years old. Starting in May 2013, i.e., the date of publication of the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), all autism disorders were merged into one umbrella diagnosis of Autism Spectrum Disorders (ASDs). Autism spectrum disorders can dramatically affect a child's life, as well as that of their families, schools, friends and a wider community. Therefore, we decided to start research on adaptation of computational intelligence methods, with particular regard to data mining and machine learning ones, for computer aided-decision support in diagnosis and therapy of persons with autism spectrum disorders (ASDs). Computer-based decision support (CDS) is defined as the use of a computer to bring relevant knowledge to bear on the health care and well-being of a patient [1]. Input data come from original author's evaluation sheets of subjects with ASDs in the important spheres (among others, self-service, communication, cognitive, physical, as well as the sphere responsible for functioning in the social and family environment, etc.). Computer-aided analysis enables us to determine trends in the abovementioned spheres (progress, stagnation, or regress) and support adjustments of the individual therapeutic and educational programs for persons covered by the care.

II. Input Data

Experiments testing the relative effectiveness of our approach have been performed on data describing over 70 cases (subjects) classified into three categories: high-functioning, medium-functioning, or low-functioning autism. Each subject has been evaluated using an original author's sheet including questions about competencies grouped into 17 spheres marked with Roman numerals:

- VI. Support for active communication.
- VII. Active communication concerning objects, people, parts of the body.
- VIII. Imitation, the length and complexity of the utterance.
- IX. Needs, emotions, moods.
- X. Object communication (the level of specific symbols).
- XI. Symbolic communication.
- XII. Requests.
- XIII. Choices.
- XIV. Communication in a pair (with contemporary, with an adult).
- XV. Social communication competences.
- XVI. Communication in a group and in social situations (in a team, in school, in the closest social environment).
- XVIII. Vocabulary.
- XIX. The degree of effectiveness of information.
- XX. The degree of motivation to communicate.
- XXI. The degree and type of hint in communication.
- XXII. Building the utterance - the degree of its complexity and functionality.
- XXIII. Dialogues.

Each case x is described by a data vector $a(x)$ consisting of over 300 descriptive attributes: $a(x) = [a_1(x), a_2(x), ..., a_m(x)]$. Such a data vector is called a profile. Four values of descriptive attributes are possible, namely 0, 25, 50, and 100. They have the following meaning:

- 0 - not performed,
- 25 - performed after physical help,
- 50 - performed after verbal help/demonstration,
- 100 - performed unaided.

If we have training data for classifiers, then to each case x we also add one decision attribute c - a class (category) to which a patient is classified. For decision attribute values, we use the following notation:

- LOW - low-functioning autism,
- $MEDIUM$ - medium-functioning autism,
- $HIGH$ - high-functioning autism.

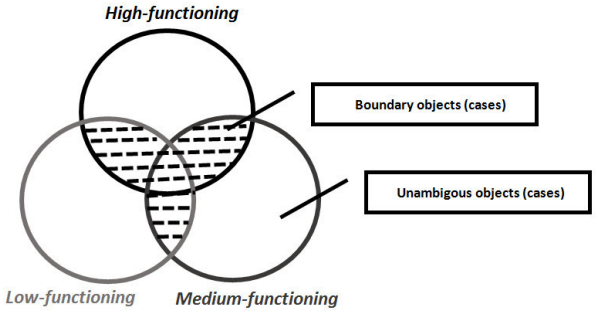


Fig. 1. Dividing a set of all training objects (cases)

In the current stage of research, each sphere is treated separately. For each sphere, the training data (which are used to learn or extract relationships between data) are stored in a tabular form (see example in Table I) which is formally called a decision table.

A decision table represents a decision system in the Pawlak's form (cf. [2]). We use the following formal definition of a decision system. A decision system DS is a tuple $DS = (U, C, D, V_{con}, V_{dec}, f_{inf}, f_{dec})$, where:

- U is a nonempty, finite set of objects,
- C is a nonempty, finite set of condition attributes,
- D is a nonempty, finite set of decision attributes,
- $V_{con} = \bigcup_{c \in C} V_c$, where V_c is a set of values of the condition attribute c,
- $V_{dec} = \bigcup_{d \in D} V_d$, where V_d is a set of values of the decision attribute d,
- $f_{inf} : C \times U \to V_{con}$ is an information function such that $f_{inf}(c, u) \in V_c$ for each $c \in C$ and $u \in U$,
- $f_{dec} : D \times U \to V_{dec}$ is a decision function such that $f_{dec}(d, u) \in V_d$ for each $d \in D$ and $u \in U$.

III. Preprocessing

Preprocessing is an important stage in data mining and knowledge discovery processes. It encompasses different tasks, e.g., extraction and selection of attributes (features), discretization of attribute values, data cleaning, etc. In this section, we describe some kind of data cleaning which is used as a preprocessing step in classification of data coming from evaluation sheets of subjects with ASDs. In our approach to classification, we can distinguish the following main stages:

1) Calculating consistency factors of objects included in the decision subsystem corresponding to class Y, with the knowledge included in the decision subsystem corresponding to class X.
2) Dividing a set of all training objects (cases) into two subsets:
   - a subset of unambiguous objects (cases),
   - a subset of boundary objects (cases).
3) Building separate classifiers trained on unambiguous objects and boundary objects, respectively.

The main aim of Stage 1 is to determine two subsets of objects included in a training data set: a subset of unambiguous objects (cases) as well as a subset of boundary objects (cases), see Figure 1.

Let $DS = (U, C, D, V_{con}, V_{dec}, f_{inf}, f_{dec})$ be a decision system, where $D = \{d\}$ and $V_{dec} = \{v_{d_1}, v_{d_2}, \ldots, v_{d_k}\}$. The set U of objects can be divided into disjoint subsets according to values of a decision attribute d, i.e.:

$$\bigcup_{i=1,2,\ldots,k} X_i,$$

where:

- $X_1 \cap X_2 \cap \cdots \cap X_k = \emptyset$,
- $X_1 \cup X_2 \cup \cdots \cup X_k = U$.

An object $u \in U$ is called a boundary object if it belongs to the subset X_i, where $i = 1, 2, \ldots, k$ and there exists X_j, where $j = 1, 2, \ldots, k$ and $j \neq i$ such that the consistency factor of u with the knowledge included in X_j is greater or equal to a given threshold θ, where $\theta \in [0, 1]$.

To differentiate two subsets of objects (unambiguous objects and boundary objects), we use an approach based on consistency factors. We assume that the boundary objects should be treated individually in a process of training the classifier (see Figure 2) because they are assigned to one decision class but they are also closed to other decision classes with respect to consistency factors. Boundary objects are intended for training more specialized and sensitive classifiers.

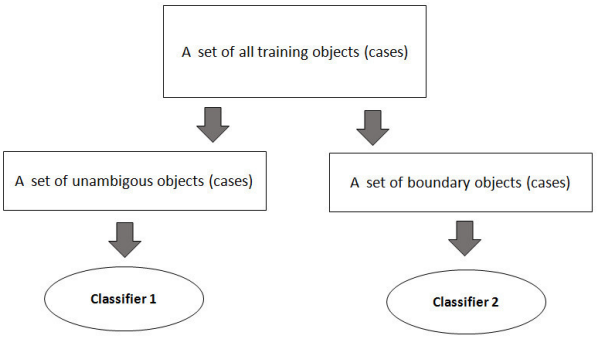


Fig. 2. Building separate classifiers

A decision system includes a finite set of cases described by attributes. Each attribute represents one of the features of cases. Apart from all cases included in the original decision system, we can consider some other cases. Such cases can be totally consistent or consistent to a certain degree with the knowledge included in the original system. The knowledge can be represented in the form of rules (production, association, etc.), cf. [3], [4]. The problem is to determine consistency factors of new cases taken into consideration with the knowledge included in the original decision system. We have adopted calculation of the consistency factor according to the definition used in [5]. That definition was derived from the approach to computing consistency factors of objects in information systems proposed in [4]. It is worth noting that an information system differs from a decision system only by the lack of decision attributes. A formal definition is as follows.

TABLE I
EXEMPLARY INPUT DATA COMING FROM THE EVALUATION SHEET

ID	VI.117	...	VI.120	VI.120a	...	VI.120f	VI.121a	...	VI.121g	VI.122	$class$
#1	50	...	100	50	...	0	50	...	0	0	LOW
...	...	...	...	...	...	...	...	...	...	...	...
#32	25	...	100	100	...	25	50	...	50	0	$MEDIUM$
...	...	...	...	...	...	...	...	...	...	...	...
#66	0	...	0	0	...	0	100	...	100	100	$HIGH$
...	...	...	...	...	...	...	...	...	...	...	...

An information system IS is a quadruple $IS = (U, A, V, f)$, where:

- U is a nonempty, finite set of objects,
- A is a nonempty, finite set of attributes,
- $V = \bigcup_{a \in A} V_a$, where V_a is a set of values of the attribute a,
- $f : A \times U \to V$ is an information function such that $f(a, u) \in V_a$ for each $a \in A$ and $u \in U$.

It is assumed, in the algorithm for computing a consistency factor, that the knowledge included in an original information system S is expressed by minimal rules true and realizable in S. Computing a consistency factor for a given object is based on determining importance (relevance) of rules extracted from the system S which are not satisfied by the new case. If the importance of these rules is greater the consistency factor of a new object with the knowledge is smaller. The importance of a set of rules not satisfied by the new case is determined by means of a strength factor of this set of rules in S. This approach has been implemented in CLAPSS (Classification and Prediction Software System) - a computer tool for solving different classification and prediction problems using, among others, some specialized approaches based mainly on the rough set theory (see [6]). The tool was designed for the Java platform. The main features of CLAPSS are the following:

- Portability. Thanks to the Java technology, the application works on various software and hardware platforms. In the future, the tool can be adapted for platforms available in mobile devices and as a service in the cloud.
- User-friendly interface (see Figure 3).
- Modularity. The project of CLAPSS and its implementation takes into consideration modularity. It makes CLAPSS possible to easily extend in the future.

Consistency factors are calculated in CLAPSS using the algorithm based on rough sets given in [7]. This algorithm makes use of important results of research on extensions of information systems given in [8]. Therefore, we recall crucial notions concerning rough sets. For more exact description and explanation we refer readers to [2] and [9].

Let $IS = (U, A, V, f)$ be an information system. Each subset $B \subseteq A$ of attributes determines an equivalence relation on U, called an indiscernibility relation $Ind(B)$, defined as

$$Ind(B) = \{(u, v) \in U \times U : \forall_{a \in B} f(a, u) = f(a, v)\}.$$

The equivalence class containing $u \in U$ will be denoted by $[u]_B$.

Let $X \subseteq U$ and $B \subseteq A$. The B-lower approximation $\underline{B}X$ of X and the B-upper approximation $\overline{B}X$ of X are defined as

$$\underline{B}X = \{u \in U : [u]_B \subseteq X\}$$

and

$$\overline{B}X = \{u \in U : [u]_B \cap X \neq \emptyset\},$$

respectively. A set $BN_B(X) = \overline{B}(X) - \underline{B}(X)$ is called the B-boundary region of X. The B-lower approximation $\underline{B}X$ of X is the set of all objects from U, which can be for certain classified as X using B, i.e., they are certainly X in view of B. The B-upper approximation $\overline{B}X$ of X is the set of all objects from U, which can be possibly classified as X using B, i.e., they are possibly X in view of B. The B-boundary region $BN_B(X)$ of X is the set of all objects from U, which can be classified neither as X nor as not-X using B. If $BN_B(X) = \emptyset$, then X is sharp (exact) with respect to B. Otherwise, X is rough (inexact).

We can provide the definition of a consistency factor (cf. [7] and [5]) in terms of appropriate lower approximations of sets. Let

- $A_{\underset{a}{\sim}} = A - \{a\}$, where $a \in A$,
- $X_a^v = \{u \in U : f(a, u) = v\}$,
- and $\widetilde{U} = \bigcup_{a \in A} \bigcup_{v \in V_a} \{\underline{A_{\underset{a}{\sim}}}(X_a^v) : \underline{A_{\underset{a}{\sim}}}(X_a^v) \neq \emptyset \wedge f(a, u^*) \neq v\}$.

The consistency factor $\xi_{IS}(u^*)$ of u^* is defined as follows:

$$\xi_{IS}(u^*) = \xi'_{IS}(u^*)\omega_{IS}(u^*),$$

where:

- $\xi'_{IS}(u^*) = 1 - \frac{card(\widetilde{U})}{card(U)}$ is a proper consistency,
- $\omega_{IS}(u^*) = \frac{card(\{a \in A : f(a,u^*) \in V_a\})}{card(A)}$ is a resemblance factor determining some affinity between the object u and objects from IS with respect to values of attributes.

A general scheme of calculating consistency factors for determining unambiguous and boundary objects is shown in Figure 4.

In experiments, for subsets of unambigous objects (cases), we have noticed significant improvement of classification accuracy (sometimes more than 10 percentage points).

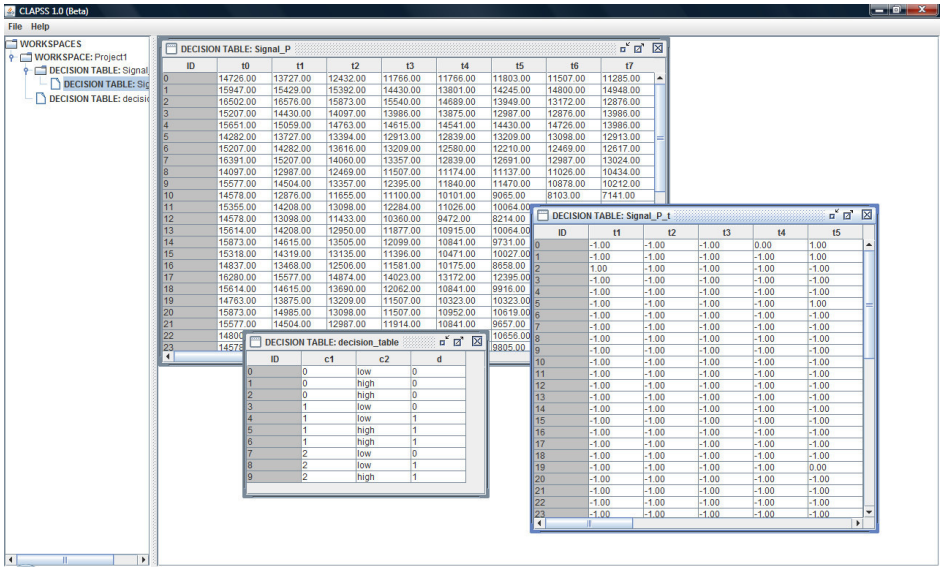


Fig. 3. User-friendly interface of CLAPSS

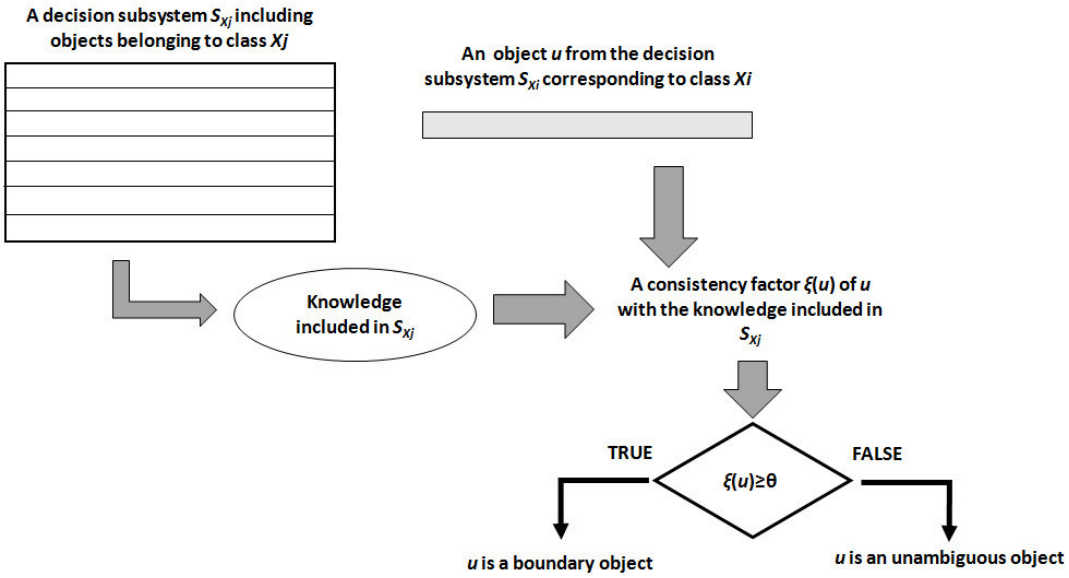


Fig. 4. Calculating consistency factors for determining unambiguous and boundary objects

IV. Conclusions and Further Work

We have described initial research on computer-aided analysis of data coming from evaluation sheets of subjects with autism spectrum disorders. This stage of research is focused on the data preprocessing step. An approach to clean a training data set for classifiers, based on consistency factors, has been proposed. The important problem in the future is to determine consistency factors of new cases taking into consideration different ways of knowledge representation. In the further stages of research, we will be interested in building hybrid classifiers combining a wide range of approaches. Adopted methods will be implemented in the specialized computer tool modelled on our previous tool, called Copernicus [10], intended for analysis and classification of data coming from the MMPI (Minnesota Multiphasic Personality Inventory) test (cf. [11]).

References

[1] R. Greenes, *Clinical Decision Support. The Road Ahead*. Elsevier Inc., 2007.

[2] Z. Pawlak, *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991.

[3] Z. Suraj, "Some remarks on extensions and restrictions of information systems," in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Artificial Intelligence, W. Ziarko and Y. Yao, Eds. Berlin Heidelberg: Springer Verlag, 2001, vol. 2005, pp. 204–211.

[4] Z. Suraj, K. Pancerz, and G. Owsiany, "On consistent and partially consistent extensions of information systems," in *Proceedings of the RSFDGrC'2005*, ser. Lecture Notes in Artificial Intelligence, D. Ślęzak *et al.*, Eds. Berlin Heidelberg: Springer Verlag, 2005, vol. 3641, pp. 224–233.

[5] Ł. Piątek, K. Pancerz, and G. Owsiany, "Validation of data categorization using extensions of information systems: Experiments on melanocytic skin lesion data," in *Proceedings of the FedCSIS'2011*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., Szczecin, Poland, 2011, pp. 147–151.

[6] K. Pancerz, "On selected functionality of the classification and prediction software system (CLAPSS)," in *Proceedings of the IDT'2015*, Zilina, Slovakia, 2015, pp. 267–274.

[7] ——, "Extensions of information systems: The rough set perspective," ser. Lecture Notes in Computer Science, J. Peters, A. Skowron, M. Chakraborty, W.-Z. Wu, and M. Wolski, Eds. Berlin Heidelberg: Springer-Verlag, 2009, vol. 5656, pp. 157–168.

[8] M. Moshkov, A. Skowron, and Z. Suraj, "On testing membership to maximal consistent extensions of information systems," in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Artificial Intelligence, S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H. S. Nguyen, and R. Slowinski, Eds. Berlin Heidelberg: Springer-Verlag, 2006, vol. 4259, pp. 85–90.

[9] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, vol. 177, pp. 3–27, 2007. doi: 10.1016/j.ins.2006.06.003

[10] K. Pancerz, O. Mich, A. Burda, and J. Gomuła, "A tool for computer-aided diagnosis of psychological disorders based on the mmpi test: an overview," in *Applications of Computational Intelligence in Biomedical Technology*, ser. Studies in Computational Intelligence, R. Bris, J. Majernik, K. Pancerz, and E. Zaitseva, Eds. Springer International Publishing, 2016, vol. 606, pp. 201–213.

[11] D. Lachar, *The MMPI: Clinical assessment and automated interpretations*. Fate Angeles: Western Psychological Services, 1974.

8th Workshop on Computational Optimization

MANY real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

We invite original contributions related to both theoretical and practical aspects of optimization methods.

Topics

The list of topics includes, but is not limited to:

- unconstrained and constrained optimization
- combinatorial optimization
- continues optimization
- global optimization
- multiobjective optimization
- optimization in dynamic and/or noisy environments
- large scale optimization
- parallel and distributed approaches in optimization
- random search algorithms, simulated annealing, tabu search and other derivative free optimization methods
- nature inspired optimization methods (evolutionary algorithms, ant colony optimization, particle swarm optimization, immune artificial systems etc)
- hybrid optimization algorithms involving natural computing techniques and other global and local optimization methods
- computational biology and optimization
- distance geometry and applications
- optimization methods for learning processes and data mining
- application of optimization methods on real life and industrial problems
- computational optimization methods in statistics, econometrics, finance, physics, chemistry, biology, medicine, engineering etc

Event Chairs

Fidanova, Stefka, Bulgarian Academy of Sciences, Bulgaria

Mucherino, Antonio, INRIA, France

Zaharie, Daniela, West University of Timisoara, Romania

Program Committee

Bartl, David, University of Ostrava, Czech Republic

Bonates, Tibérius, Universidade Federal do Ceará, Brazil

Breaban, Mihaela

Chira, Camelia

Fidanova, Stefka, Bulgarian Academy of Science

Gonçalves, Douglas, Universidade Federal de Santa Catarina, Brazil

Gualandi, Stefano

Hosobe, Hiroshi, Hosei University, Japan

Iiduka, Hideaki, Kyushu Institute of Technology, Japan

Krislock, Nathan, Northern Illinois University, United States

Lavor, Carlile, IMECC-UNICAMP, Brazil

Marinov, Pencho, Bulgarian Academy of Science, Bulgaria

Mihalas, Stelian, West University of Timisoara

Muscalagiu, Ionel, Politehnica University Timisoara, Romania

Nannicini , Giacomo

Ninin, Jordan, ENSTA-Bretagne, France

Parsopoulos, Konstantinos, University of Patras

Pintea, Camelia, Tehnical University Cluj-Napoca, Romania

Pop, Petrica

Roeva, Olympia, Institute of Biophysics and Biomedical Engineering, Bulgaria

Siarry, Patrick, Universite Paris XII Val de Marne, France

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

Stefanov, Stefan, South-West University ""Neofit Rilski, Bulgaria

Stuetzle, Thomas, Université Libre de Bruxelles (ULB), Belgium

Suganthan, Ponnuthurai Nagaratnam, Nanyang Technological University, Singapore

Tamir, Tami, The Interdisciplinary Center (IDC), Israel

Tvrdik, Josef, University of Ostrava, Czech Republic

Voller, Zach

Vrahatis, Michael, University of Patras, Greece

Zilinskas, Antanas, Vilnius University, Lithuania

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 71–78
DOI: 10.15439/2015F311
ACSIS, Vol. 6

Time-Dependent Traveling Salesman Problem with Multiple Time Windows

Jarosław Hurkała
Institute of Control & Computation Engineering, Warsaw University of Technology, Warsaw, Poland
Interdisciplinary Center for Security, Reliability and Trust – University of Luxembourg
Email: J.Hurkala@elka.pw.edu.pl

***Abstract*—The TSP, VRP and OP problems with time constraints have one common sub-problem – the task of finding the minimum route duration for a given order of customers. While much work has been done on routing and scheduling problems with time windows, to this date only few articles considered problems with multiple time windows. Moreover, since the assumption of constant travel time between two locations at all times is very unrealistic, problems with time-dependent travel were introduced and studied. Finally, it is also possible to imagine some situations, in which the service time changes during the day. Again, both issues have been investigated only in conjunction with single time windows. In this paper we propose a novel algorithm for computing minimum route duration in traveling salesman problem with multiple time windows and time-dependent travel and service time. The algorithm can be applied to wide range of problems in which a traveler has to visit a set of customers or locations within specified time windows taking into account the traffic and variable service/visit time. Furthermore, we compare three metaheuristics for computing a one-day schedule for this problem, and show that it can be solved very efficiently.**

I. Introduction

In this paper we focus our work on time-dependent routing and scheduling problem with multiple time windows. The problem consist of an agent (tourist, sales representative, etc.) whose aim/duty is to visit a predefined set of customers/locations (e.g. points of interest). Each customer/location may define many time windows which indicates the availability during the day. Furthermore, we assume, that the travel time between the customers/locations changes due to the traffic. In this work we also assume different service/visit time in different time windows.

In this work we describe the theory and algorithms for computing one-day schedule in time-dependent traveling salesman problem with multiple time windows for application in many well known operational research problems such as vehicle routing problem (VRP) (see [2], [3], [8], [14]), orienteering problem (OP) (see [15], [16]), or generally traveling salesman problem (TSP) with complex time constraints. While much work has been done on mixed routing and scheduling problems with time windows, to this date only few articles considered problems with multiple time windows (cf. [2]).

Throughout this article, we will denote a sequence of customers as a route, while we use the term schedule to denote a route with fixed visit times.

The paper is organized as follows: in Section 2 we describe the problem, and discuss additional issues arising from multiple time windows, and time-dependent travel and service time. Section 3 describes preprocessing of time windows and presents the minimum route duration algorithm. In Section 4 we explain in details the algorithms used for computing the one-day schedule in time-dependent traveling salesman problem with multiple time windows. Section 5 shows the results of our numerical experiments. Finally, some concluding remarks are given in Section 6.

II. Problem description

We consider a Time-Dependent Traveling Salesman Problem with Multiple Time Windows (TDTSPMTW) with the following features:

1) each customer can define multiple time windows during which he is available and can be serviced;
2) the service time can be different in every time window of the customer;
3) the travel time depends on the traffic time zone, in which the transit actually occurs;
4) starting and ending depots are treated as customers so that they also have time windows.

The TDTSPMTW problem can be defined as follows.

A. Problem notation

Let $\mathcal{I} = \{1, \ldots, n\}$ be the set of customers $i \in \mathcal{I}$ that are visited by the traveling salesman. Let $\mathcal{W}_i$, be the set of i-th customer time windows $j \in \mathcal{W}_i$, during which the visit can take place. The set of time windows of all the customers will be denoted by $\mathcal{W} = \bigcup_{i \in \mathcal{I}} \mathcal{W}_i$. Thus, $[a_i^j, b_i^j]$ will denote the j-th time window of i-th customer, where a_i^j is the beginning, and b_i^j is the end of the time window, and the service time of i-th customer in j-th time window will be denoted by s_i^j. Let $\mathcal{Z}$, $k \in \mathcal{Z}$, be the set of traffic time zones $[p^k, q^k]$, where p^k is the beginning, and q^k is the end of the traffic time zone, and t_i^k be the travel time from i-th customer to the next one in the sequence, in k-th time zone. *Notice, that we deliberately define travel to the next customer instead of to the current one - this significantly simplifies the considerations*. The visit at i-th customer will be denoted by $[\alpha_i, \beta_i]$, where α_i is the time of arrival at the customer, and β_i is the departure time of the visit.

B. Traffic time zones

Before we can proceed with explaining the minimum route duration algorithm, the problem of traffic time zones has to be accommodated. Since the customers can already have multiple time windows, we can take advantage of this property and create additional, „virtual“ time windows so that the travel time in each window is well-defined. For every time window we have to check whether it lies in one traffic time zone, or maybe spans across multiple zones. In the latter case, the original time window has to be divided into smaller, overlapping windows. We shall explain this on the following example. Let $[a, b]$ be a time window with service time s, that spreads over two time zones: $[p^1, q^1]$ with travel time t^1 and $[p^2, q^2]$ with travel time t^2, such that $p^1 < a < q^1 = p^2 < b < q^2$. Than, we need to divide the original window into the following ones: $[a, q^1]$ with service time s and travel time t^1 and $[p^2 - s, b]$ with service time s and travel time t^2. Notice, that the beginning of the second window is brought forward by the service time, because we consider travel time to the next customer in sequence, and the transit starts as soon as the current customer has been serviced. Thus, let $\mathcal{V} = \bigcup_{i \in \mathcal{I}} \mathcal{V}_i$ be the set of virtual time windows of customers $i \in \mathcal{I}$. Finally, we can define a function $\gamma : \mathcal{V} \to \mathcal{Z}$ that maps the virtual time windows into the time zones. Since each virtual time window falls within one traffic time zone, we know that the function is well-defined.

C. The normalized formulation

Having the travel time unambiguously defined for every time window we can normalize the model similarly to [8], [16]. Thus, we merge the service time and the travel time into one parameter $d_i^j = s_i^j + t_i^{\gamma(j)}$, denoting the visit duration. At the same time, we postpone the ending of every (virtual) time window by the travel time associated with it, i.e. windows $[a_i^j, b_i^j]$ are transformed into $[a_i^j, b_i^j + t_i^{\gamma(j)}]$. *Notice, that from this point on in the article the departure time will have new meaning, i.e., the moment the salesman reaches the next customer in the sequence.*

D. Master problem

Let $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ be the permutation of customers.

The master problem of TDTSPMTW, i.e. the problem of finding optimal sequence of customers to be visited during one day, can be defined as follows:

$$\pi^* = \arg\min_{\pi} MinimumRouteDuration(\pi) \tag{1}$$

The problem of finding the minimum route duration for a given sequence of customers (i.e. the subproblem of TDTSPMTW) can be formulated as follows.

E. Subproblem formulation

Let us define the main decision variables and explain their meaning. For the time windows selection we define:

$$y_i^j = \begin{cases} 1 & \text{if the visit at customer i takes place} \\ & \text{within the time window j} \\ 0 & \text{else} \end{cases} \tag{2}$$

Using these notations, we write a mathematical model for the TDTSPMTW subproblem. The objective is to minimize the route duration (calculated as a difference between departure time of last customer and arrival time at first customer):

$$\min \beta_n - \alpha_1 \tag{3}$$

$$(a_i^j + d_i^j) y_i^j \le \beta_i \le b_i^j + \mathcal{M}(1 - y_i^j) \quad i \in \mathcal{I}, j \in \mathcal{V} \tag{4}$$

$$\sum_{j \in \mathcal{V}_i} y_i^j = 1 \quad i \in \mathcal{I} \tag{5}$$

$$\alpha_i = \beta_i - \sum_{j \in \mathcal{V}_i} d_i^j y_i^j \quad i \in \mathcal{I} \tag{6}$$

$$\beta_i \le \alpha_{i+1} \quad i \in \mathcal{I} \setminus \{n\} \tag{7}$$

$$y \in \{0, 1\}, \beta \ge 0 \tag{8}$$

Constraints (4) handle the time windows in a classical way (the departure time must be within a time window) with the noticeable addition of the upper index, since we have multiple time windows. Constraints (5) ensure that exactly one time window per customer is chosen. The auxiliary constraints (6) compute the start of every visit (arrival time α). Finally, constraints (7) forbid to start the next visit before the current customer departure time, so that the visits do not overlap. The y variables are binary, and departure times β are non-negative real variables (8).

III. Minimum route duration algorithm

The minimum route duration algorithm that we have developed requires a feasible solutions to start with. Hence, the preprocessing of the virtual time windows is needed.

A. Preprocessing

In order for the algorithm to work, unnecessary time windows from the bottom-right (see Algorithm 1) and top-left (see Algorithm 2) corners have to be removed (if you imagine the first window of the first customer in the bottom-left corner of a diagram, and the last window of the last customer in the top-right corner). Similar approach has already been proposed in [16], but in our formulation the visit duration may be different in subsequent time windows of a customer, hence we can dismiss only the ones that lead to unfeasible schedule.

Algorithm 1 TDTSPMTW: top-down preprocessing

```
1: α ← max_{j∈V_n} {b_n^j − d_n^j}
2: for i = n − 1 to 1 do
3:     V_i ← V_i \ {j ∈ V_i : a_i^j + d_i^j > α}
4:     if V_i = {∅} then
5:         return false {schedule not feasible}
6:     end if
7:     for j ∈ V_i do
8:         b_i^j ← min{b_i^j, α}
9:     end for
10:    α ← max_{j∈V_i} {b_i^j − d_i^j}
11: end for
12: return true
```

Algorithm 2 TDTSPMTW: bottom-up preprocessing

1: $\beta \leftarrow \min_{j \in \mathcal{V}_1}\{a_1^j + d_1^j\}$
2: **for** $i = 2$ **to** n **do**
3: $\quad \mathcal{V}_i \leftarrow \mathcal{V}_i \setminus \{j \in \mathcal{V}_i : b_i^j - d_i^j < \beta\}$
4: $\quad$ **if** $\mathcal{V}_i = \{\emptyset\}$ **then**
5: $\quad\quad$ **return false** {schedule not feasible}
6: $\quad$ **end if**
7: $\quad$ **for** $j \in \mathcal{V}_i$ **do**
8: $\quad\quad a_i^j \leftarrow \max\{a_i^j, \beta\}$
9: $\quad$ **end for**
10: $\quad \beta \leftarrow \min_{j \in \mathcal{V}_i}\{a_i^j + d_i^j\}$
11: **end for**
12: **return true**

Algorithm 3 TDTSPMTW: constructing feasible sub-schedule

Require: i, β
1: **while** $i \leq n$ **do**
2: $\quad \beta \leftarrow \min_{j \in \mathcal{V}_i, b_i^j - d_i^j \geq \beta} \max\{a_i^j, \beta\} + d_i^j$
3: $\quad i \leftarrow i + 1$
4: **end while**
5: **return** β

Algorithm 4 TDTSPMTW: constructing dominant sub-schedule

Require: i, α
1: **while** $i \geq 1$ **do**
2: $\quad \alpha \leftarrow \max_{j \in \mathcal{V}_i, a_i^j + d_i^j \leq \alpha} \min\{b_i^j, \alpha\} - d_i^j$
3: $\quad i \leftarrow i - 1$
4: **end while**
5: **return** α

Algorithm 5 TDTSPMTW: minimum route duration

1: $\tau^* \leftarrow \infty$
2: **if** TopDownPreprocessing() **and** BottomUpPreprocessing() **then**
3: $\quad$ **for** $i = 1$ **to** n **do**
4: $\quad\quad$ **for** $j \in \mathcal{V}_i$ **do**
5: $\quad\quad\quad \beta \leftarrow constructFeasibleSubSchedule(i + 1, b_i^j)$
6: $\quad\quad\quad \alpha \leftarrow constructDominantSubSchedule(i - 1, b_i^j - d_i^j)$
7: $\quad\quad\quad \tau \leftarrow \beta - \alpha$
8: $\quad\quad\quad$ **if** $\tau < \tau^*$ **then**
9: $\quad\quad\quad\quad \tau^* \leftarrow \tau$
10: $\quad\quad\quad$ **end if**
11: $\quad\quad$ **end for**
12: $\quad$ **end for**
13: **end if**
14: **return** τ^*

The top-down preprocessing algorithm (Algorithm 1) begins with calculating the latest possible arriving at the last customer (denoted by α). Than, starting from the second to last customer, virtual time windows are removed, for which the visit starting at the beginning of the window would exceed α (line 3). If after this step there is no more time windows, the schedule (sequence of customers) is not feasible, and the algorithm terminates. Otherwise, the remaining windows are tightened so that the ending of each window does not exceed α (line 8). Finally, the α value is recalculated for the current customer (line 10) and the procedure starts over with previous customer.

The bottom-up preprocessing algorithm (Algorithm 2) works almost identically as top-down. The differences are as follows:

- the earliest possible departure time from the customer (denoted by β) is taken into consideration (lines: 1, 10);
- algorithm iterates from the second customer to the last one;
- windows are removed if the beginning of visit that starts as late as possible overlaps the departure time β (line 3);
- the remaining windows are tightened so that they do not begin earlier than β (line 8).

The minimum route duration algorithm that we propose in this paper consists of iteratively reviewing schedules of which the one with the shortest duration is chosen. The procedure of constructing each schedule is divided into two phases.

B. Phase 1: constructing feasible sub-schedule

The first phase consist of computing a feasible sub-schedule that starts from a given customer, and ends with the last one (see Algorithm 3). This procedure requires an index of a customer and an initial departure time. Repeatedly, among the virtual time windows that have enough room to fit the visit after the given departure time ($b_i^j - d_i^j \geq \beta$), the earliest visit departure time (calculated as: $\max\{a_i^j, \beta\} + d_i^j$) is searched for.

C. Phase 2: constructing dominant sub-schedule

This phase is based on the notion of the dominant solutions (see [8], [14], [16]). To put it simply, a schedule with starting time α^1 dominates schedule with starting time α^2, if $\alpha^1 > \alpha^2$ and at the same time ending time $\beta^1 = \beta^2$ (cf. [8]). In our procedure, instead of fixed visit departure time, we use the arrival time (obviously, $\alpha = \beta - d$). Nevertheless, the principle is the same - we repeatedly search for the latest possible starting time of a visit (calculated as: $\min\{b_i^j, \alpha\} - d_i^j$) among the virtual time windows, that are suitable ($a_i^j + d_i^j \leq \alpha$), i.e. windows in which the currently considered visit does not overlap the initial one.

D. The main algorithm

The algorithm enumerates schedules, constructing them in a particular fashion. For each virtual time window, first, a feasible sub-schedule is constructed with the initial departure time set to the **end** of the current time window (line 5). Secondly, a dominant sub-schedule is constructed with the initial arrival time set to the latest possible start of the visit in the considered window (line 6). The route duration is than computed as the difference between the departure time from

last customer and arrival time at the first customer (line 7). The best solution found during the process is stored (lines 8–10) and returned (line 14). For the algorithm overview see Algorithm 5.

Although the main procedure of the algorithm looks self-explanatory, the reason it finds the optimal solution is not trivial. Savelsbergh [14] has introduced the concept of forward time slack to postpone the beginning of service at a given customer. It can be proven, that the optimal schedule can be postponed until one of the visits **ends** with the time window. Hence, by iteratively reviewing schedules one by one so that every possible time window with visit at the end of it is taken into consideration, an optimal schedule is found by our algorithm.

E. Computational complexity

The preprocessing procedures have both $O(|\mathcal{V}|)$ time complexity. The sub-schedule construction procedures have together $O(|\mathcal{V}|)$ time complexity (they consider disjoint sets of time windows). The main procedure has $O(|\mathcal{V}|)$ time complexity (every time window is taken into consideration). Hence, the total time complexity of the algorithm is $O(|\mathcal{V}|^2)$.

IV. TDTSPMTW MASTER PROBLEM ALGORITHMS

For solving the TDTSPMTW master problem, i.e. computing the one-day schedule, we have chosen three different metaheuristics.

A. Simulated annealing

Simulated annealing was first introduced by Kirkpatrick [10], while Černý [1] pointed out the analogy between the annealing process of solids and solving combinatorial problems. Researchers have been studying the application of the SA algorithm in various fields of optimization. Koulamas [11] presented a survey of operational research problems in which the heuristic was applied. The effectives of the algorithm was also inspected in particular by Hurkała and Hurkała [5], [6], and also Hurkała and Śliwiński [7].

The optimization process of the simulated annealing algorithm can be described in the following steps. Before the algorithm can start, an initial solution is required. Then, repeatedly, a candidate solution is randomly chosen from the neighborhood of the current solution. If the candidate solution is the same or better than the current one, it is accepted and replaces the current solution. A worse solution than the current one still has a chance to be accepted with, so called, acceptance probability. This probability is a function of difference between objective values of both solutions and depends on a control parameter taken from the thermodynamics, called temperature. The temperature is decreased after a number of iterations, and the process continues as described above. The optimization is stopped either after a maximum number of iterations or when a minimum temperature is reached. The best solution found during the annealing process is considered final. For the algorithm overview see Algorithm 6.

Algorithm 6 Simulated annealing

Require: Initial solution s_1
1: $s^* \leftarrow s_1$
2: **for** $i = 1$ **to** N **do**
3: **for** $t = 1$ **to** N_{const} **do**
4: $s_2 \leftarrow perturbate(s_1)$
5: $\delta \leftarrow C(s_2) - C(s_1)$
6: **if** $\delta \leq 0$ or $e^{-\delta/k\tau} > random(0,1)$ **then**
7: $s_1 \leftarrow s_2$
8: **end if**
9: **if** $C(s_2) < C(s^*)$ **then**
10: $s^* \leftarrow s_2$
11: **end if**
12: **end for**
13: $\tau \leftarrow \tau * \alpha$
14: **end for**
15: **return** s^*

The main building block of the simulated annealing is the temperature decrease (also known as the cooling process), which consists of decreasing the temperature by a reduce factor. The parameters associated with this mechanism are as follows:

1) Initial temperature.
2) Function of temperature decrease in consecutive iterations.
3) The number of iterations at each temperature (Metropolis equilibrium).
4) Minimum temperature at which the algorithm terminates or alternatively the maximum number of iterations as the stopping criterion.

Let τ be the temperature and α be the reduce factor. Then the annealing scheme can be represented as the following recursive function:

$$\tau^{i+1} = \alpha * \tau^i, \tag{9}$$

where i is the number of current iteration in which the cooling schedule takes place.

Second building block of SA that has to be customized for a particular problem is the acceptance probability function, which determines whether to accept or reject candidate solution that is worse than the current one. The most widely used function is the following:

$$p\left(\delta, \tau\right) = e^{-\delta/k\tau}, \tag{10}$$

where $\delta = E(s_2) - E(s_1)$ is the difference between the objective value (denoted by E) of the candidate (s_2) and the current solution (s_1), and k is the Boltzmann constant found by:

$$k = \frac{\delta^0}{\log \frac{p^0}{\tau^0}}, \tag{11}$$

where δ^0 is an estimated difference between objective values of two solutions, p^0 is the initial value of the acceptance probability and τ^0 is the initial temperature. Notice that we use

decimal logarithm rather than natural, which is most widely seen in the literature.

B. List-based threshold accepting

List-based threshold accepting algorithm (LBTA) introduced by Lee [12], [13] is an extent of the threshold accepting meta-heuristic, which belongs to the randomized search class of algorithms. The search trajectory crosses the solution space by moving from one solution to a random neighbor of that solution, and so on. Unlike the greedy local search methods which consist of choosing a better solution from the neighborhood of the current solution until such can be found (hill climbing), the threshold accepting allows choosing a worse candidate solution based on a threshold value. In the general concept of the threshold accepting algorithm it is assumed that a set of decreasing threshold values is given before the computation or an initial threshold value and a decrease schedule is specified. The rate at which the values decrease controls the trade-off between diversification (associated with large threshold values) and intensification (small threshold values) of the search. It is immensely difficult to predict how the algorithm will behave when a certain decrease rate is applied for a given problem without running the actual computation. It is also very common that the algorithm with the same parameters works better for some problem instances and significantly worse for others. These reflections led to the list-based threshold accepting branch of threshold accepting meta-heuristic.

In the list-based threshold accepting approach, instead of a predefined set of values, a list is dynamically created during a presolve phase of the algorithm. The list, which in a way contains knowledge about the search space of the underlying problem, is then used to solve it.

The first phase of the algorithm consists of gathering information about the search space of the problem that is to be solved. From an initial solution a neighbor solution is created using a move function (perturbation operator) chosen at random from a predefined set of functions. If the candidate solution is better than the current one, it is accepted and becomes the current solution. Otherwise, a threshold value is calculated as a relative change between the two solutions:

$$\Delta = (C(s_2) - C(s_1))/C(s_1) \tag{12}$$

and added to the list, where $C(s_i)$ is the objective function value of the solution $s_i \in S$, and S is a set of all feasible solutions. For this formula to work, it is silently assumed that $C : S \to \mathbb{R}_+ \cup \{0\}$. This procedure is repeated until the specified size of the list is reached. For the algorithm overview see Algorithm 7.

The second phase of the algorithm is the main optimization routine, in which a solution to the problem is found. The algorithm itself is very similar to that of the previous phase. We start from an initial solution, create new solution from the neighborhood of current one using one of the move function, and compare both solutions. If the candidate solution is better, it becomes the current one. Otherwise a relative change is calculated. To this point algorithms in both phases are identical. The difference in the optimization procedure is that we compare the threshold value with the largest value from the list. If the new threshold value is larger, then the new solution is discarded. Otherwise, the new threshold value replaces the value from the list, and the candidate solution is accepted to next iteration. The best solution found during the optimization process is considered final.

Algorithm 7 Creating the list of threshold values

Require: Initial solution s_1, list size S, set of move operators $m \in M$

```
i ← 0
while i < N do
    m ← random(M)
    s2 ← m(s1)
    if C(s1) ≤ C(s2) then
        Δ ← (C(s2) − C(s1))/C(s1)
        list ← list ∪ {Δ}
        i ← i + 1
    else
        s1 ← s2
    end if
end while
return list
```

The list-based threshold accepting algorithm also incorporates early termination mechanism: after a (specified) number of candidate solutions is subsequently discarded, the optimization is stopped, and the best solution found so far is returned. The optimization procedure of the list-based threshold accepting algorithm is shown in Algorithm 8.

The original LBTA algorithm does not have a solution space independent stopping criterion. If the number of subsequently discarded worse solutions is set too high, the algorithm will run for an unacceptable long time (it has been observed during preliminary tests). Hence, we propose to use additionally a global counter of iterations so that when a limit is reached, the algorithm terminates gracefully.

In the first phase of the list-based threshold accepting algorithm the list is populated with values of relative change between two solutions $\Delta \geq 0$. After careful consideration, we believe that including zeros in the list is a misconception. In the actual optimization procedure, i.e. the second phase, the threshold value is computed only if the new solution is worse than the current one, which means that the calculated relative change will always have a positive value ($\Delta_{new} > 0$). The new threshold value is compared with the largest value from the list (T_{hmax}). Thus, we can distinguish three cases:

1) $T_{hmax} = 0$: since thresholds are non-negative from definition, in this case the list contains all zero elements and it will not change throughout the whole procedure (T_{hmax} is constant). Comparing a positive threshold value Δ_{new} against zero yields in discarding the candidate solution. The conclusions are as follows:
    a) it does not matter how many zeros are in the list,

Algorithm 8 LBTA optimization procedure

Require: Initial solution s_1, thresholds list L, set of move operators $m \in M$

$i \leftarrow 0$
$s^* \leftarrow s_1$
while $i \leq N$ **do**
  $m \leftarrow random(M)$
  $s_2 \leftarrow m(s_1)$
  $i \leftarrow i + 1$
  if $C(s_2) \leq C(s_1)$ **then**
    if $C(s_2) \leq C(s^*)$ **then**
      $s^* \leftarrow s_2$
    end if
    $s_1 \leftarrow s_2$
    $i = 0$
  else
    $\Delta_{new} \leftarrow (C(s_2) - C(s_1))/C(s_1)$
    if $\Delta_{new} < max(list)$ **then**
      $list \leftarrow list \setminus \{max(list)\}$
      $list \leftarrow list \cup \{\Delta_{new}\}$
      $s_1 \leftarrow s_2$
      $i = 0$
    end if
  end if
end while
return s^*

the effective size of the list is equal to one,

b) the algorithm is reduced to hill climbing algorithm that accepts candidate solutions which are at least as good as the current one.

2) $T_{hmax} > 0$ and $\Delta_{new} < T_{hmax}$: the largest (positive) threshold value from the list T_{hmax} is replaced by a smaller (positive) threshold value Δ_{new}. The number of zero elements in the list remains the same throughout the whole procedure and therefore is completely irrelevant to the optimization process. The effective list size is equal to the number of positive elements.
3) $T_{hmax} > 0$ and $\Delta_{new} \geq T_{hmax}$: the new solution is discarded and the list remains unchanged.

The main idea behind the list is to control the diversification and intensification of the search process. In the early stage of the search the algorithm should allow to cover as much solution space as possible, which means that the thresholds in the list are expected to be large enough to make that happen. In the middle stage, the algorithm should slowly stop fostering the diversification and begin to foster the intensification of the search. In the end stage, the intensification should be the strongest, i.e. the list is supposed to contain smaller and smaller threshold values, which induces discarding of worse solution candidates. As a consequence, the algorithm is converging to a local or possibly even a global optimum.

Algorithm 9 Variable neighborhood descent

Require: Initial solution s_0

$s_1 \leftarrow s_0$
$i \leftarrow 1$
repeat
  for $j = 1$ **to** $size(N_i)$ **do**
    $s_2 \leftarrow N_i(s_0)$
    if $C(s_2) < C(s_1)$ **then**
      $s_1 \leftarrow s_2$
    end if
  end for
  if $C(s_1) < C(s_0)$ **then**
    $s_0 \leftarrow s_1$
    $i \leftarrow 1$
  else
    $i \leftarrow i + 1$
  end if
until $i = |N|$
return s_0

Algorithm 10 Reduced variable neighborhood search

Require: Initial solution s_0

$i \leftarrow 1$
repeat
  $s_1 \leftarrow VND(N_i(s_0))$
  if $C(s_1) < C(s_0)$ **then**
    $s_0 \leftarrow s_1$
    $i \leftarrow 1$
  else
    $i \leftarrow i + 1$
  end if
until $i = |N|$
return s_0

C. Variable neighborhood search

Variable neighborhood search (VNS) is a metaheuristic algorithm proposed by Mledanović [9]. This global optimization method is based on an idea of systematically changing the neighborhood in the descent to local minima and in the escape from the valleys which contain them. It has already been successfully used in different Vehicle Routing Problems.

The VNS algorithm consists of two building blocks: variable neighborhood descent (VND) and reduced variable neighborhood search (RVNS).

The optimization process of VND can be explained as follows. First, an initial solution is required. Within a given neighborhood a candidate solution is repeatedly generated and it replaces the current one if it is better. After a specified number of iterations (neighborhood size) the neighborhood is changed to the first one if a better than initial solution has been found. Furthermore, the best solution found so far replaces the initial solution. Otherwise, if the search resulted in no better solutions that the initial one, the current neighborhood is changed to the next one. Either way, the whole operation is

repeated again until the search gets stuck in a local optimum. The best solution found during this process is returned. For the overview of the algorithm see Algorithm 9.

The RVNS is a stochastic algorithm that executes the VND with different initial solutions. This simple procedure, which in fact is quite similar to the VND, can be described in the following few steps. First, an initial solution - a starting point - is required. Within a given neighborhood a candidate solution is repeatedly generated from the initial one, and passed to the VND procedure. If the VND returns a solution that is better than the current one, it gets replaced, and the algorithm starts over from the first neighborhood. Otherwise, the neighborhood is changed to the next one. The whole process can be repeated until the stopping criterion (e.g. specified number of evaluations, time limit) is met. The optimization procedure of RVNS is shown in Algorithm 10.

D. Neighborhood function

The most problem-specific mechanism of SA, LBTA and VNS algorithms, that always needs a different approach and implementation, is the procedure of generating a candidate solution from the neighborhood of the current one, which is called a perturbation scheme, transition operation/operator or a move function. Although there are many ways to accomplish this task for the traveling salesman problem, we have chosen the following three operators:

1) interchanging two adjacent customers,
2) interchanging two random customers,
3) moving a single, random customer to a randomly chosen position.

V. Numerical experiments

The numerical experiments were performed on a number of randomly generated problem instances of different size. The algorithms were implemented in C++. All the computations were executed on the Intel Core i7 3.4GHz microprocessor.

To better compare relative performance of the three algorithms, the only stopping criterion for single run was reaching the same number of schedule evaluations for all computations and problem sizes. This way we could compare the speed as well as the convergence per iteration.

The resulting one-day schedules are presented in Table I. The first column indicates the instance number. The number of customers (second column) ranges from 13 to 23. The route durations found by the three algorithms are shown in columns 3-5. The values in columns 6-8 indicate relative difference between the algorithms outcomes. In order of brevity, we show the computation time in one (the last) column for the given problem instance - it was almost the same for every algorithm due to the (identical) stopping criterion.

The algorithms produced similar results in terms of both the solution quality and the computation time. For some instances one algorithm produces better results, while for some other it is the other way around. Generally, the VNS tends to find a little bit better solutions: 2.49% on average than LBTA, and 2.97% than SA. LBTA and SA are on the other hand almost identical - the former is on average better by only 0.49% than the latter.

VI. Conclusions

We have developed a novel and efficient algorithm that computes the minimum route duration for the Time-Dependent Traveling Salesman Problem with Multiple Time Windows, and compared three metaheuristic algorithms that computes the one-day schedule, which can be successfully utilized in time-oriented TSP, VRP, OP, and other mixed routing and scheduling problems. The minimum route duration algorithm guarantees finding optimal solution in quadratic time in terms of the total number of time windows. To our knowledge, this is the first attempt of solving this kind of time-dependent TSP with multiple time windows.

Acknowledgment

This research was financed by the European Union through the European Regional Development Fund under the Operational Programme “Innovative Economy” for the years 2007-2013; Priority 1 – Research and development of modern technologies under the project POIG.01.03.01-14-076/12: “Decision Support System for Large-Scale Periodic Vehicle Routing and Scheduling Problems with Complex Constraints” and has been supported by the European Union in the framework of European Social Fund through the project: Supporting Educational Initiatives of the Warsaw University of Technology in Teaching and Skill Improvement Training in the Area of Teleinformatics.

References

[1] Černý, V., Thermodynamical approach to traveling salesman problem: An efficient simulation algorithm. J. Optim. Theory Appl., vol. 45 (1985) 41–51.

[2] Belhaiza S., P. Hansen, G. Laporte. *A hybrid variable neighborhood tabu search heuristic for the vehicle routing problem with multiple time windows*, Computers & Operations Research, Volume 52 (2014), 269-281.

[3] Favaretto D., E. Moretti, P. Pellegrini. *Ant colony system for a VRP with multiple time windows and multiple visits*, Journal of Interdisciplinary Mathematics, Volume 10, Issue 2 (2007), 263-284.

[4] Hurkała J., *Minimum Route Duration Algorithm for Traveling Salesman Problem with Multiple Time Windows*, Vehicle Routing and Logistics Optimization, June 8-10, 2015, Vienna, Austria [accepted for presentation].

[5] Hurkała, J. and Hurkała, A., Effective Design of the Simulated Annealing Algorithm for the Flowshop Problem with Minimum Makespan Criterion, Journal of Telecommunications and Information Technology 2 (2012) 92–98.

[6] Hurkała, J. and Hurkała, A., Fair optimization with advanced aggregation operators in a multicriteria facility layout problem, in Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, IEEE, 2013, 355–362.

[7] Hurkała, J. and Śliwiński, T., Fair flow optimization with advanced aggregation operators in Wireless Mesh Networks, in Proceedings of the Federated Conference on Computer Science and Information Systems, IEEE, 2012, 415–421.

[8] Jong C., G. Kant, A. van Vliet. *On Finding Minimal Route Duration in the Vehicle Routing Problem with Multiple Time Windows*, Tech. Rep., The Netherlands: Department of Computer Science, Utrecht University, 1996.

[9] Mladenović, N., and Hansen, P. *Variable neighborhood search*. Computers and Operations Research 24 (11), 1997, 1097-?1100.

[10] Kirkpatrick, S., Gellat, C.D. and Vecchi, M.P., Optimization by simulated annealing, Science, vol. 220 (1983) 671–680.

TABLE I
ROUTE DURATIONS AND COMPUTATION TIME

#	$\|\mathcal{I}\|$	VNS	LBTA	SA	$VNS/LBTA$	$VNS/LBTA$	$LBTA/SA$	Time [s]
1	13	28113	28532	29608	-1.47%	-5.05%	-3.63%	0.281
2	13	27921	27921	29793	0.00%	-6.28%	-6.28%	0.314
3	13	31193	31193	31193	0.00%	0.00%	0.00%	0.283
4	13	29913	31540	31591	-5.16%	-5.31%	-0.16%	0.293
5	13	28113	28532	29608	-1.47%	-5.05%	-3.63%	0.224
6	13	27921	27921	29793	0.00%	-6.28%	-6.28%	0.264
7	14	28314	27299	28289	3.72%	0.09%	-3.50%	0.299
8	15	28134	28039	28325	0.34%	-0.67%	-1.01%	0.381
9	15	28134	28039	28325	0.34%	-0.67%	-1.01%	0.243
10	16	30143	31400	30442	-4.00%	-0.98%	3.15%	0.376
11	16	27822	30210	30078	-7.90%	-7.50%	0.44%	0.282
12	16	35438	35438	35438	0.00%	0.00%	0.00%	0.405
13	16	29957	31453	29843	-4.76%	0.38%	5.39%	0.298
14	17	32442	33306	33306	-2.59%	-2.59%	0.00%	0.499
15	17	33609	33609	33609	0.00%	0.00%	0.00%	0.444
16	17	31585	31730	32515	-0.46%	-2.86%	-2.41%	0.285
17	17	30691	30450	31648	0.79%	-3.02%	-3.79%	0.279
18	17	29136	29136	30553	0.00%	-4.64%	-4.64%	0.295
19	17	31882	34707	34707	-8.14%	-8.14%	0.00%	0.401
20	18	36246	37569	36759	-3.52%	-1.40%	2.20%	0.517
21	18	36389	38696	38696	-5.96%	-5.96%	0.00%	0.399
22	19	40102	40102	40102	0.00%	0.00%	0.00%	0.390
23	19	36231	36231	36231	0.00%	0.00%	0.00%	0.388
24	20	37982	40187	39746	-5.49%	-4.44%	1.11%	0.571
25	20	38107	39943	39897	-4.60%	-4.49%	0.12%	0.360
26	20	38511	38195	38163	0.83%	0.91%	0.08%	0.525
27	20	38081	35533	38577	7.17%	-1.29%	-7.89%	0.470
28	20	40163	40707	40707	-1.34%	-1.34%	0.00%	0.341
29	20	37977	38930	38930	-2.45%	-2.45%	0.00%	0.303
30	21	38468	40434	40379	-4.86%	-4.73%	0.14%	0.510
31	21	39072	39855	39711	-1.96%	-1.61%	0.36%	0.449
32	21	36876	40170	38456	-8.20%	-4.11%	4.46%	0.423
33	21	34468	37702	37702	-8.58%	-8.58%	0.00%	0.376
34	21	39069	40912	40912	-4.50%	-4.50%	0.00%	0.363
35	21	39857	40497	40497	-1.58%	-1.58%	0.00%	0.347
36	22	38856	40044	39298	-2.97%	-1.12%	1.90%	0.550
37	22	37854	40243	40243	-5.94%	-5.94%	0.00%	0.554
38	22	40608	40608	40608	0.00%	0.00%	0.00%	0.557
39	22	40017	40017	40017	0.00%	0.00%	0.00%	0.522
40	22	39997	39997	39997	0.00%	0.00%	0.00%	0.495
41	22	38292	40830	40830	-6.22%	-6.22%	0.00%	0.380
42	22	36783	40711	39449	-9.65%	-6.76%	3.20%	0.446
43	23	36057	36057	37616	0.00%	-4.14%	-4.14%	0.725
44	23	37725	40688	40688	-7.28%	-7.28%	0.00%	0.533
45	23	38952	40366	39465	-3.50%	-1.30%	2.28%	0.538

[11] Koulamas, C., Antony, S.R. and Jaen, R., A survey of simulated annealing applications to operations research problems, Omega, 22 (1994) 41–56.

[12] Lee, D.S., Vassiliadis, V.S., Park, J.M., A novel thresholdaccepting metaheuristic for the job-shop scheduling problem. Computers & Operations Research, 31 (2004) 2199–2213.

[13] Lee, D.S., Vassiliadis, V.S., Park, J.M., List-Based Threshold-Accepting Algorithm for Zero-Wait Scheduling of Multiproduct Batch Plants, Ind. Eng. Chem. Res. 41 (25), 2002, pp. 6579-?6588.

[14] Savelsbergh, M.W.P. *The Vehicle Routing Problem with Time Windows: Minimizing Route Duration*, ORSA Journal on Computing, Vol. 4, Issue 2 (1992), 146-161.

[15] Souffriau W., P. Vansteenwegen, G.V. Berghe, D. Van Oudheusden. . *The Multi-Constraint Team Orienteering Problem with Multiple Time Windows*, Transportation Science, Volume 47, Issue 1 (2013), 53-63.

[16] Tricoire F., Romauch, M., Doerner, K.F., Hartl, R.F. *Heuristics for the multi-period orienteering problem with multiple time windows*, Computers & Operations Research 37 (2010), 351-367.

Computer Science & Systems

CSS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- BCPC'15—1st International Workshop on Biological, Chemical and Physical Computations
- CANA'15—8th Computer Aspects of Numerical Algorithms
- IWCPS'15—2nd International Workshop on Cyber-Physical Systems
- MMAP'15—8th International Symposium on Multimedia Applications and Processing
- WAPL'15—5th Workshop on Advances on Programming Languages

1st International Workshop on Biological, Chemical and Physical Computations

THE First International Workshop on Biological, Chemical and Physical Computations (BCPC'15) is a unique interdisciplinary workshop which brings together computer scientists and engineers with biologists, chemists, and physicists to initiate development of novel paradigms, architectures and implementations of computing devices adopting principles of information processing in physical, chemical and biological systems. The workshop aims to bring together world-leading scientists whose research focuses on non-traditional theoretical machines, experimental prototypes and genuine implementations of non-classical computing devices, who try to revisit existing approaches in unconventional computing, provide scientists and engineers with blueprints of realizable computing devices, and take a critical glance at the design of novel and emergent computing systems to point out failures and shortcomings of both theoretical and experimental approaches.

Topics

The topics of interest include, but are not limited to:

- Physics of computation,
- Slime mould computing,
- Social insects computing,
- Chemical computing,
- Bio-molecular computing,
- Cellular automata as models of massively parallel computing,
- Logics of unconventional computing,
- Reaction-diffusion computing,
- Molecular machines incorporating information processing,
- Memristors,
- Organic electronics,
- Noise-based computing,
- Novel hardware systems,
- Mechanical computing,
- Physical limits to mechanical computation.

Event Chairs

Adamatzky, Andrew, Professor, UWE, Bristol, UK, United Kingdom

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

Schumann, Andrew, University of Information Technology and Management in Rzeszow, Poland

Program Committee

Akl, Selim, Queen's School of Computing, Canada

Armstrong, Rachel, Architecture, Planning and Landscape, University of Newcastle, United Kingdom

Asai, Tetsuya, Laboratory of Advanced LSI Engineering, Hokkaido University, Japan

Costa, Jose-Felix, Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Costello, Ben De Lacy, Centre for Research in Analytical, Material and Sensor Sciences, University of the West of England, United Kingdom

Di Ventra, Max, Department of Physics, University of California, United States

Dittrich, Peter, Bio Systems Analysis Group, Friedrich-Schiller-University Jena, Germany

Erokhin, Victor, Departament of Phisics, University of Parma, Italy

Gorecki, Jerzy, Institute of Physical Chemistry, PAN, Poland

Hanczyc, Martin, Institute of Physics Chemistry and Pharmacy, University of Southern Denmark, Denmark

Miranda, Eduardo, University of Plymouth, United Kingdom

Petre, Ion, Computer Science Programme at Åbo Akademi University, Finland

Shirakawa, Tomohiro, National Defense Academy of Japan, Japan

Sirakoulis, Georgios, Department of Electrical & Computer Engineering, Democritus University of Thrace, Greece

Stannett, Mike, University of Sheffield, United Kingdom

Stepney, Susan, University of York, United Kingdom

Tsuda, Soichiro, School of Chemisty, University of Glasgow, United Kingdom

Tucker, John, Department of Computer Science, Swansea University, United Kingdom

Zauner, Klaus-Peter, Faculty of Physical Sciences and Engineering, University of Southampton, United Kingdom

Zelinka, Ivan, Department of Computer Science, VSB Technical University in Ostrava, Czech Republic

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 83–87

DOI: 10.15439/2015F391
ACSIS, Vol. 6

Predicting Metal-Binding Sites of Protein Residues

Serkan R. Küçükbay, Hasan Oğul
Department of Computer Engineering,
Başkent University, Ankara, Turkey
Email: skucukbay@gmail.com,
hogul@baskent.edu.tr

Abstract—**Metal ions in protein are critical to the function, structure and stability of protein. For this reason accurate prediction of metal binding sites in protein is very important. Here, we present our study which is performed for predicting metal binding sites for histidines (HIS) and cysteines from protein sequence. Three different methods are applied for this task: Support Vector Machine (SVM), Naive Bayes and Variable-length Markov chain. All these methods use only sequence information to classify a residue as metal binding or not. Several feature sets are employed to evaluate impact on prediction results. We predict metal binding sites for mentioned amino acids at 35% precision and 75% recall with Naive Bayes, at 25% precision and 23% recall with Support Vector Machine and at 0.05% precision and 60% recall with Variable-length Markov chain. We observe significant differences in performance depending on the selected feature set. The results show that Naive Bayes is competitive for metal binding site detection.**

I. Introduction

Protein plays a crucial role in all biological processes. *And* they consist of one or more long chains of amino acid residues. In the frame of this perspective, amino acids are important ligands with nitrogen and oxygen as the donor, constituent of many biological important molecules [1].

It is estimated that approximately half of all *proteins* contain a metal [2]. A significant fraction (about one third) of all known proteins is believed to bind metal ions as cofactors in their native conformation [3]. The biological activities of proteins require these cofactors to assist their daily routines. For this reason, a metal ion in a protein and prediction of its binding point is very important to understand the function of proteins in biological activities. Metal ions in proteins are responsible for multiple tasks. They help stabilizing protein structure [4], induce conformational changes [5–7], and assist protein functions (e.g. electron transfer, nucleophilic catalysis).

There are many related studies about predicting metal binding sites, however, machine learning techniques have been recently applied to predict the metal binding sites of residues.

Predicting metal binding sites by using non-computational methods has some drawbacks. X-ray absorption spectroscopy (HT-XAS) has been recently proved to be capable of identifying metalloproteins with high reliability [8, 9]. However, the specific ligands involved in binding metal ion(s) cannot be identified by these techniques [9]. Motif-based system has also been developed by using regular expressions but since regular expressions can be quite specific, their results have many false negatives. To overcome these drawbacks, many computational learning techniques have been developed to predict metal binding sites. Early approaches can be found in the work of Nakata et al. (1995). In this study, they focused on predicting zinc-finger DNA-binding proteins with a neural network. In this approach, applicable results were generated by a method for certain types of zinc-binding protein in spite of limitation about scarcity of data at that time. Recently-developed approaches for metal-binding sites prediction have mainly focused on CYS only [10], CYS and HIS binding transition metals [3] or CYS, HIS, ASP, and GLU binding zinc ions [12, 13].

In addition, recent studies in predicting metal binding sites indicate that Support Vector Machine is a popular machine learning technique in this area. In many works, Support Vector Machine was employed as a single solution of a problem or it was used with some other techniques to predict metal binding sites. For example, developed architecture consists of two stages. In the first stage of this study, Support Vector Machine was employed for local classification and these outputs were used as inputs for second stage to refine these local predictions [13].

In this study, we employed three different methods to predict metal binding for CYS and HIS by using only sequence information and amino acid composition: Support Vector Machine, Naive Bayes and Variable-length Markov Chain. Obtained results were compared with each other to give information for future works. Furthermore, we used some different feature compositions to train our model and prediction results were compared to give some clues about used features which were valuable for metal binding sites prediction.

This paper is organized as follows: in Section 2, we provide detailed description of materials and methods. Our obtained results are discussed in Section 3. We finally draw some conclusions in Section 4.

II. Materials and Methods

A. Feature Representation

To obtain feature vector for SVM and Naïve Bayes, we found all target residues (CYS and HIS) for each protein sequence. Each target residue refers individual instances for our input vectors. These individual instances were obtained by building a window. This window consisted of flanking residues of target residue and target residue was located in the middle of the window. Each position in the window was represented by the corresponding value of PAM 120 (Table I. values of pam are shown) substitution matrix. By this way, each position of window was represented by 23 finite numbers. In other words, the total size of window equals to (2w-1) X 23 because the target residue value was ignored for input vectors. Also some characteristic values of selected window were obtained by running script version of Protein Composition Server [14] and these values were added to input vectors. Additionally, we used four different types of amino acid compositions. These are Pseudo Amino Acid Composition (**PAAC**), Amphiphilic Pseudo Amino Acid Composition (**APAAC**), Five Factor Score of Amino Acid Composition (**5FSS**) and PhysioChemical Properties (**PC**). By using different types of amino acid compositions, we made four different individual instances for each target residues. Hence, we could compare their contributions to prediction results. Finally, one additional feature was coded for the relative position of the target residue with respect to the sequence length for each individual instance. Input vector generation steps are illustrated in Fig 1.

To obtain features for Variable-length Markov chain, first, we found all target residues for each protein sequence. Then, we built a window which consisted of flanking residues of the target residue and this target residue was located in the middle of the window. Window size was set to 15. After building the window, it was split at its center. After splitting process, four different feature sets were created. This feature sets contained different parts of the window according to the metal binding state of target residue. Content of these feature sets differed from each other. The mentioned content was constructed via the following steps.

If target residue is metal-bonded and flanking amino acids are located at the left side of target residue, one part of the window is considered as a left-side metal-bonded sample.

If target residue is metal-bonded and flanking amino acids are located at the right side of target residue, one part of the window is considered as a right-side metal-bonded sample. If target residue is not metal-bonded and flanking amino acids are located at the right side of target residue, one part of the window is considered as a right-side non-metal bonded sample. If target residue is not metal-bonded and flanking amino acids are located at the left side of target residue, one part of the window is considered as a left-side non-metal bonded sample.

B. Classification

Support Vector Machines: In this study, we applied three different methods for our problem. One of these is Support Vector Machine (SVM).

SVM learning has been applied to many prediction problems in bioinformatics. Furthermore, obtained prediction results are acceptable for many problems in this search area. For this reason, we chose SVM for our solution. For classification, publicly-available at github libsvm package was used to implement SVM. The radial basis kernel was selected for this process (gamma = 0.05, cost = 0.1).

In this process, SVM was used to train feature vectors of

TABLE I.
PAM 120 MATRIX

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-8
N	-1	-1	4	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	5	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5	-4	3	0	-3	-4	-3	-3	-2	1	-4	-3	-2	-8
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	5	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	8	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	8	-5	-3	-4	-1	4	-3	-5	-6	-3	-8
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-3	-2	-3	-5	6	1	-1	-7	-6	-2	-2	-1	-2	-8
S	1	-1	1	0	0	-2	-1	1	-2	-2	-4	-1	-2	-3	1	3	2	-2	-3	-2	0	-1	-1	-8
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	4	-6	-3	0	0	-2	-1	-8
W	-7	1	-4	-8	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-6	12	-2	-8	-6	-7	-5	-8
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	8	-3	-3	-5	-3	-8
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	5	-3	-3	-1	-8
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	-3	4	2	-1	-8
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	4	-1	-8
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-3	-2	-1	-1	-5	-3	-1	-1	-1	-2	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

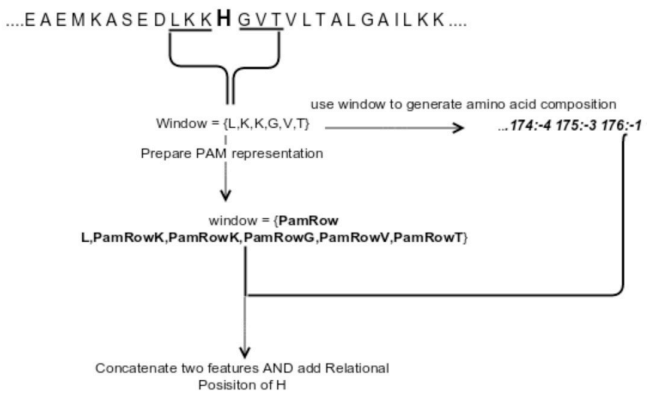


Fig.1 Input vector generation steps

the training data set and to make the prediction on the test data set. Detailed information about data set is mentioned in Section 2.3.

Naïve Bayes: As mentioned above, we applied different methods on our test sets and another one of these methods is Naive Bayes. The Naïve Bayes classifier has proved to be very effective in many real data applications [17]. Naive Bayes classifiers are of the family of simple probabilistic classifiers. It is based on applying Bayes' theorem with strong independence assumptions between the features.

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(v-\mu_c)^2}{2\sigma_c^2}} \quad (1)$$

For classification, the publicly-available MATLAB Naive Bayes package was used to train and predict our data set. We chose Gaussian Naïve Bayes because our data set consisted of continuous data. Applied kernel function is shown in Equation 1.

C. A Generative Approach: Variable Length Markov Chain

Markov chains are used to model sequential data in terms of the order of individual letters. In zero-order Markov Chain, the likelihood of a sequence S_1^N is given by the probability that is obtained by multiplying the probabilities of each symbols contained, i.e.,

$$P(S_1^N) = \prod_{j=1}^{N} P(S_j = s_j) \quad (2)$$

where P(.) refers to probability, S_j is the random variable representing the letter position j with s_j as its realization.

A more flexible version of higher order Markov models allows a variable length that depends on the preceding subsequence to given position such that the order of the model becomes a function the context at each position. This model is called as Variable Length Markov Chain (VLMC) built on the sequence likelihood defined as:

$$P(S_1^N) = \prod_{j=1}^{N} P\left(S_j = s_j \mid S_{j-L_j}^{j-1} = s_{j-L_j}^{j-1}\right) \quad (3)$$

where L_j is the optimal length preceding subsequences respectively and $S_{j-L_j}^{j-1}$ is that sub-sequences.

An efficient implementation of VLMC can be realized using Probabilistic Suffix Trees (PST). The PST method was introduced by Bejerano and Yona to model the protein families [15]. The original PST model was based on identifying significant short segments among the many input sequences, regardless of the relative position of these segments within the different proteins [16]. In this study, to classify a sequence into one of the families, a separate PST is constructed for each family in the data set, and according to the probability distribution over PST, a probability that the sequence belongs to that family is assigned to the query sequence. By comparing this probability score the sequence is determined as belonging to that family or not.

For this approach, we created four different train data sets for training processes as mentioned in feature representation section. We trained each data set to obtain probabilistic suffix trees(PST) so we created four different PST (PST1 consists of data such as flanking amino acids that are located at the left side of metal-bonded CYS or HIS; PST2 consists of data such as flanking amino acids that are located at the left side of CYS or HIS which are not bonded by a metal; PST3 consists of data such as flanking amino acids that are located at the right side of metal-bonded CYS or HIS; PST4 consists of data such as flanking amino acids that are located at the right side of CYS or HIS which are not bonded by a metal.) for each train data set. After PST generation, we built a window for each CYS and HIS from test sets. Then, for each created window, we ran prediction processes for all obtained PSTs. Finally, the outputs of the prediction processes were compared with each other and we marked predicted CYS or HIS as metal bonded or not by evaluating comparison results. However, before comparison, we multiplied outputs of metal-bonded and nonmetal-bonded ones between each other.

D. Dataset

We used a non-redundant set of PDB containing 2727 protein sequences to test our methods. The used data set was prepared by [3] for their research. The detailed and well defined explanation can be found in the mentioned paper. In Table II, we listed some information about this data set.

Table III.
NUMBER OF CYS & HIS AND THEIR STATE OF METAL BOUNDED

	Metal Bounded	Non-Metal Bounded
CYS	933	4702
HIS	678	12982

TABLE IVII.
CHANGE OF SVM PREDICTION RESULTS ACCORDING TO SELECTED FEATURE FOR TRAINING

PAM	Relative Position	5FSS	APAAC	PAAC	PC	RECALL	PRECISION	AUC
X	-	-	-	-	-	0.22	0.24	0.58
X	-	X	-	-	-	0.23	0.25	0.58
X	X	X	-	-	-	0.22	0.24	0.59
X	-	-	X	-	-	0.11	0.18	0.51
X	X	-	X	-	-	0.11	0.18	0.61
X	-	-	-	X	-	0.24	0.27	0.60
X	X	-	-	X	-	0.24	0.27	0.45
X	-	-	-	-	X	0.11	0.14	0.45
X	X	-	-	-	X	0.11	0.13	0.45
X	X	-	-	-	-	0.23	0.25	0.60

TABLE IIIV
CHANGE OF NAIVE BAYES PREDICTION RESULTS ACCORDING TO SELECTED FEATURE

PAM	Relative Position	5FSS	APAAC	PAAC	PC	RECALL	PRECISION	AUC
X	-	-	-	-	-	0.65	0.45	0.78
X	-	X	-	-	-	0.75	0.35	0.80
X	X	X	-	-	-	0.72	0.36	0.76
X	-	-	-	X	-	0.65	0.44	0.77
X	X	-	-	X	-	0.66	0.45	0.77
X	-	-	-	-	X	0.51	0.11	0.59
X	X	-	-	-	X	0.50	0.13	0.64
X	-	-	X	-	-	0.43	0.18	0.62
X	X	-	X	-	-	0.43	0.18	0.61

TABLE V
PREDICTION RESULT OF VARIABLE LENGTH MARKOV CHAIN

PRECISION	RECALL	AUC
0.05	0.60	0.39

III. EVALUATION CRITERIA

In this work, we use precision, recall and area under the curve as performance measurements. The precision was defined as **TP/(TP + FP)**, where **TP** (true positives) was referred to the number of correctly-identified positive examples (metal binding residues); **FP** (false positive) was the number of negative examples (residues predicted to bind metal, although they do not bind to a metal according to PDB) that were incorrectly predicted as positive. The recall was defined as **TP/(TP + FN)**, where **FN** (false negative) was the number of positive examples that were incorrectly predicted as negative. In this study, the negative examples were far more than the positive examples. For such an unbalanced dataset, Area Under Curve (AUC) can present an overly optimistic view of the performance of a method. To

obtain AUC values, we used publicly available MATLAB package.

IV. Results

In this study, we created ten different feature vectors to train with SVM and Naive Bayes. Also we evaluated the predictions for Variable-length Markov chain. All obtained scores are listed in Table III, Table IV and Table V.

The obtained results show us that Naive Bayes is competitive for metal binding site detection.

On the other hand, we used varied feature combinations and they give us a chance to evaluate their prediction score changes according to feature type. For ex; using pam matrix representation is very smart way to identify amino acid for classification because the result is really acceptable and length of this feature limited by number of amino acid count in nature. Also using global descriptors as a feature is practicable for this area.

As a result, we presented a method to predict metal binding sites from amino acid sequences by SVM, Naive Bayes and Variable-length Markov chain .We obtained many results for different feature sets and we reached higher results with Naive Bayes(used features were PAM and 5FSS). The mentioned case predicted CYS/HIS with 35% precision at 75% recall level and 80% AUC value, when tested on a non-redundant set of PDB containing 2727 unique protein chains.

V. Conclusion

Predicting metal-binding conformations of proteins through computational techniques is a favorable effort in the wake of estimating final protein structures. In this study, we evaluate different feature representation schemes and implement different methods to predict metal binding sites of protein residues. Obtained results are compared with each other and valuable feature types are observed. The results justify that Naive Bayes approach can produce acceptable predictions for residue classification. We believe that this study is going to lead our future works and our approach can have an impact on metal binding site detection. We will use Naive Bayes classification for large data set using big data technologies such as spark and storm.

Acknowledgments

We would like to thank those who publicly share their developed codes, scripts and experimental data and those who maintain these useful sharings.

References

[1] J. Reedijk, “Comprehensive Coordination Chemistry”, vol. 2, chp. 13.2, Pergamon, Oxford, pp. 73-98, 1987.

[2] A. J. Thomson and H. B. Gray “Bio-inorganic chemistry”, Current Opinion in Chemical Biology 2: 155–158.

[3] A. Passerini, M. Punta, A. Ceroni, B. Rost, and P Frasconi,“Identifying Cysteines and Histidines in Transition-Metal-Binding Sites Using Support Vector Machines and Neural Networks,” Proteins, vol. 65, no. 2, pp. 305-316, 2006.

[4] L. Bancini et. al., “A prokaryotic superoxide dismutase paralog lacking two Cu ligands: from largely unstructured in solution to ordered in the crystal”, Proc Natl Acad Sci USA, 102:7541-7546, 2005.

[5] M. Akke, T. Drakenberg and WJ. Chazin, “Three-dimensional solution structure of Ca(2+)-loaded porcine calbindin D9k determined by nuclear magnetic resonance spectroscopy”, 31:1011-1020, 1992.

[6] H. M. Greenblatt, H. Feinberg, PA. Tucker and G. Shoham, “Carboxypeptidase A: native, zinc-removed and mercury-replaced forms”, 54:289–305, 1998.

[7] H. Sun, H. Li and PJ. Sadler, “Transferrin as a metal ion mediator”, Chem Rev., 99: 2817-2842, 1999.

[8] M. R. Chance and W. Shi, “Metallomics and metalloproteomics.”, Cell Mol. Life Sci., 65, 3040-3048,2008.

[9] W. Shi et. al., “Characterization of metalloproteins by high-throughput X-ray absorption spectroscopy”, Genom Res., 21(6):898-907, 2011.

[10] A. Passerini, M. Lippi and P. Frasconi, “MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence”, Nucleic Acids Res., 39(Web Server issue):W288-92, 2011.

[11] F. Ferre and P. Clote, “DiANNA 1.1: An Extension of the DiANNA Web Server for Ternary Cysteine Classification”, Nucleic Acids Research, vol.34, pp.W182-W185, 2006.

[12] A. Passerini, C. Andreini, S. Menchetti, A. Rosato, and P. Frasconi, “Predicting Zinc Binding at the Proteome Level,” BMC Bioinformatics, vol. 8, p. 39, 2007.

[13] N. Shu, T. Zhou, and S. Hovmoller, “Prediction of Zinc-Binding Sites in Proteins from Sequence,” Bioinformatics, vol. 24, no. 6, pp. 775-782, 2008.

[14] L. Rishishwar, N. Mishra, B. Pant, K. Pant, and K. R. Pardasani, ProCoS - PROtein COmposition Server, Bioinformation, 5(5): 227. PMC: 3040505, 2010.

[15] G. Bejenora and G. Yona, “Variations on probabilistic suffix trees: statistical modelling and prediction of protein families”, Bioinformatics Vol.17 No.1, pp. 23-43, 2000.

[16] H. Oğul and E. Mumcuoğlu, “SVM-based detection of distant protein structural relationships using pairwise probabilistic suffix trees”, Computational Biology and Chemistry Vol.30, pp. 292-299, 2006.

[17] M. Boulle, “Parsimonious Naïve Bayes”, 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 355-359, 2014.

8th Workshop on Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

Topics

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on GPUs
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

Event Chairs

Bylina, Jaroslaw, Maria Curie-Sklodowska University, Poland

Bylina, Beata, Maria Curie-Sklodowska University, Poland

Stpiczyński, Przemysław, Maria Curie-Sklodowska University, Poland

Program Committee

Amodio, Pierluigi, Università di Bari, Italy

Anastassi, Zacharias, Qatar University, Qatar

Banaś, Krzysztof, AGH University of Science and Technology, Poland

Barán, Benjamín, Universidad Nacional del Este

Brugnano, Luigi, Universita' di Firenze, Italy

Czachorski, Tadeusz, IITiS

Filippone, Salvatore, University Rome Tor Vergata, Italy

Filote, Constantin

Fourneau, Jean-Michel

Gansterer, Wilfried, University of Vienna, Austria

Georgiev, Krassimir, IICT - BAS, Bulgaria

Gimenez, Domingo, University of Murcia, Spain

Gravvanis, George, Democritus University of Thrace, Greece

Knottenbelt, William, Imperial College London, United Kingdom

Kozielski, Stanislaw

Kucaba-Pietal, Anna, Politechnika Rzeszowska, Poland

Lirkov, Ivan, Institute of Information and Communication Technologies , Bulgarian Academy of Sciences, Bulgaria

Maksimov, Vyacheslav, Institute of Mathematics and Mechanics, Russia

Marowka, Ami, Bar-Ilan University, Israel

Petcu, Dana, West University of Timisoara, Romania

Pultarova, Ivana, Czech Technical University in Prague, Czech Republic

Satco, Bianca-Renata, Stefan cel Mare University of Suceava, Romania

Sedukhin, Stanislav, The University of Aizu, Japan

Sergeichuk, Vladimir, Institute of Mathematics of NAS of Ukraine, Ukraine

Shishkina, Olga, Max Planck Institute for Dynamics and Self-Organization, Germany

Srinivasan, Natesan, Indian Institute of Technology, India

Szadkowski, Zbigniew, University of Lodz, Poland

Szajowski, Krzysztof, Institute of Mathematics and Computer Science, Poland

Trivedi, Kishor S., Duke University, United States

Tudruj, Marek, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland

Tůma, Miroslav, Academy of Sciences of the Czech Republic, Czech Republic

Ustimenko, Vasyl, Marie Curie-Sklodowska University, Poland

Vazhenin, Alexander, University of Aizu, Japan

Wójcik, Grzegorz M., Institute of Computer Science, Maria Curie-Sklodowska University, Poland

Wyrzykowski, Roman, Czestochowa University of Technology, Poland

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 91–96

DOI: 10.15439/2015F233
ACSIS, Vol. 6

jPar – a simple, free and lightweight tool for parallelizing Matlab calculations on multicores and in clusters

Andrzej Karbowski
NASK, Research and Academic
Computer Network
ul. Wąwozowa 18
02-796 Warszawa, Poland
and
Institute of Control
and Computation Engineering
Warsaw University of Technology
ul. Nowowiejska 15/19
00-665 Warszawa, Poland
E-mail: A.Karbowski@elka.pw.edu.pl

Marek Majchrowski
Warsaw University of Technology
E-mail: M.Majchrowski@coi.pw.edu.pl

Piotr Trojanek
Institute of Control
and Computation Engineering
Warsaw University of Technology
E-mail: P.Trojanek@elka.pw.edu.pl

Tomasz Pokorski
Institute of Control
and Computation Engineering
Warsaw University of Technology
E-mail: T.Pokorski@stud.elka.pw.edu.pl

Dawid Załuga
Institute of Control
and Computation Engineering
Warsaw University of Technology
E-mail: D.Zaluga@stud.elka.pw.edu.pl

***Abstract*—We present a very simple, free tool for parallelizing calculations under Matlab in multicore and cluster environments. After the installation it does not use any compilers, MEX files, disk files, etc. It is compatible with the old Paralize package, but allows the involved cores/machines to do other jobs when a worker core/machine is not busy.**

I. Introduction

THE aim of the described work was to add to the Matlab software a free and lightweight support for distributed and parallel computations. The main idea was to provide a simple, user friendly tool (such as Matlab environment itself), which does not use too much the machines' resources to organize the calculations.

Despite the existence of Matlab Parallel Computing Toolbox (PCT) and Distributed Computing Server (DCS) [1] there is still some sense in developing such tools, because:

- Matlab PCT/DCS are changing, that is, there are some differences between the subsequent versions. For example, in recent versions there were changes concerning two commands: `matlabpool` (which was replaced in R2013b release by `parpool` and completely removed since release R2015a) and `mapreduce`,
- if one wants to perform computations in clusters, except Matlab PCT it is necessary to buy Matlab DCS, which is much more expensive and complicated in administration,
- universities and research laboratories usually buy Matlab licenses in packs; this means, that after hours in labs many licenses are free,
- there is a big interest in free software tools.

Due to these reasons, we observe still a lot of works devoted to developing software environments to parallelize Matlab, both on multicores and in clusters [2], to mention only: *Multicore* [3], *MatlabMPI* [4], *pMatlab* [5] and *shared-matrix* [6].

Almost all free packages for parallelizing Matlab calculations use disk files for communication between Matlab instances, what is inefficient and prone to errors. They are often based on C-MEX files, which have to be ported to every environment (i.e., compiler, operating system), where the tool will be used. Some of them also make use of the remote execution of the child processes through the shells such as `ssh`, `rsh`, which are neither convenient to use, nor available everywhere (e.g., in Windows operating system). Most of these packages introduces plenty of Matlab functions or extend basic syntax to provide communication and synchronization between computing nodes. And last, but not least, these packages do not always work under new versions of Matlab.

Our goal was to find a solution for the described issues and to provide a smart, but easy to install and use tool for parallel and distributed computing, working under different versions of Matlab.

One of the oldest and - no doubts - the simplest software packages to parallelize Matlab calculations is *Paralize* [7]. It consists of only two m-functions: `paralize.m` and `serve.m` of, respectively, 146 and 94 lines (including comment lines) and implements fork-join model of parallel computations, without communication and synchronization between instances. It helps to make calculations both within multiprocessor/multicore machines and in clusters of computers. The biggest drawback of *Paralize* is that communication and synchronization is realized via the disk files, what involves active polling, or, in other words, busy waiting and sometimes causes race conditions errors.

The *jPar* package is as simple as *Paralize* from the user point of view (actually it is compatible with it), very easy to install, but it does not waste the cycles of cores on active polling and allows to use the machines with started Matlab instances to other purposes. Moreover, it does not use for communication and synchronization network filesystem, but more efficient and reliable Java RMI and Java threads.

II. JPAR PROJECT AND IMPLEMENTATION

The basic requirements for *jPar* were:

- to avoid communication via disk files, which is slow and may cause race conditions errors,
- to avoid active polling, that is the waste of time of processors/cores (and energy),
- it should be portable (implemented in Java and .m scripts),
- it should be possibly small,
- it should be easy to install and use.

Distribution of the work in *jPar* is done by dividing the data and executing the same operation on all chunks by slave Matlab processes called *solvers*. The data chunk together with a function to be executed (*task*) is passed to *solvers*. When the calculations are finished, the results are gathered on the *console* (i.e., the *client*) Matlab session (Fig. 1), so the calculations with *jPar* may be interpreted as a simplified implementation of the distributed arrays idea.

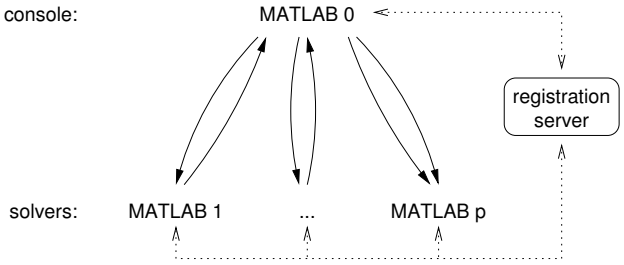


Fig. 1. *jPar* communication model.

The presented approach has been implemented using Java language, owing to its portability and ease of interfacing with Matlab [8]. Synchronization and data exchange between the nodes were done by means of Java *RMI* (Remote Method Invocation) mechanism.

The package consists of only three components:

1) Registration server
2) Solvers
3) Client.

The first process to be run is the *registration server*, which task is to manage the set of *solvers*. It is provided in the form of a single Java executable JAR file (which also contains all the Java classes used by Matlab) and should be started from the command line of the operating system.

The next step is starting several Matlab instances: one for the *jPar console* and the others for *solvers*. The latter should be started from within Matlab sessions (from the same directory) and left. It does not matter what operating systems are used, since both computing (Matlab) and communication (Java) environments are system independent. The registration is done by adding a handle to a remote *solver* object to the managed set. Then the Matlab sessions are blocked until the new tasks are available. The *client* just divides input data into chunks (as in *Paralize*, along the third dimension) and creates partial tasks to perform the distributed job.

In our first version *solver* was implemented as a non-blocking function. Calling Matlab from Java was done by using *JMI* (*Java Matlab Interface*) with functions and classes defined in the Matlab JMI package (JAR file JMI.jar). However, this implementation required running *solvers* within Matlab sessions with GUI display. Another important drawback of this solution was, that the JMI package was distributed by MathWorks "as it is", without any warranty, support or documentation.

Hence, in the final version *solver* is a blocking Matlab function, which waits on a Java object, until it is notified. Then, using provided methods, *solver* gathers all parameters, converts them from Java representation and makes calculations in Matlab. At the end the results are converted and sent back to the *jPar client*.

As it was mentioned before, to perform the computation on multiple nodes, the input data has to be in the form of a three dimensional array. The job is divided into chunks by the partition along the third coordinate, which identifies the task (and the chunk). All the remaining parameters are passed unmodified to *solvers*.

In the example presented below ten *tasks* will be created (each to handle 100×100 matrix):

```
>> a=rand(100,100,10)+i*rand(100,100,10);
>> [V,D]=jpar_client('eig',a);
```

Marshalling and unmarshalling the data objects is done by RMI, but some attention was also paid to the transfer of data between Matlab and Java. While basic types (such as `double`) are automatically converted, the imaginary part of a number is discarded. There is also an issue with vectors, where dimension has to be preserved, since Java does not distinct between horizontal and vertical arrangement of the elements. In these cases Matlab data is converted by the package to the internal representation:

```
public class JMArray
  implements Serializable {

  private Object realpart, imagpart;
  private int dimX, dimY;
  /* ... */
}
```

Since the options of some *solvers* are passed as Matlab structures (e.g., in *fmincon* function from Matlab Optimization Toolbox), the package also converts them to an internal representation:

```
public class JMStruct
```

```
  implements Serializable {

  private Object fields, values;
  private int dimX, dimY;
  /* ... */
}
```

The variables *dimX* and *dimY* are also used to reshape the structure after Matlab→Java→Matlab conversion to match the original arrangement. In *fields* the package sustains the field names of the original Matlab structure, *values* are collected as in the original structure.

It is possible to pass strings or function handles as second and further arguments of *j*Par, that is the actual parameters of parallelized functions. To make it possible, a Java class *JMHandle* in package *matlab.jmhandle* was created. The purpose of this class is to store both the information about the name of the function used to create handle and the absolute path to the file containing that function. When we are passing a function handle to *j*Par, an object - an instance of *JMHandle* class - is created. When the *JMHandle* object is received, a Matlab function handle is recreated basing on the information stored in the object.

It is also possible to pass sparse matrices as arguments. To pass a single sparse matrix as an argument of a function being parallelized no special effort is needed. In order to pass multiple sparse matrices designated for parallelization, a vector of cells containing these sparse matrices must be created and passed as an argument to the function being parallelized. Sparse matrices must be stored in a cell array of size $1 \times N$, where N is the number of matrices. All matrices in this "vector" must be sparse.

For example:

```
>> as=sparse(a);
>> bs=sparse(b);
>> cs=sparse(c);
>> con{1} = as;
>> con{2} = bs;
>> con{3} = cs;
```

When a proper argument is passed, *j*Par converts sparse matrices to vectors that represent coordinates and values. These vectors are passed to *JMSparse* class in Java. After that the data is transferred between the *client* and a *solver*. Before the computations, the *solver* is restoring sparse matrices from the vectors. If results of function being parallelized are sparse matrices, *j*Par puts them into the vector of cells.

The complete implementation of *jPar* has about 800 lines of Java code and 400 lines of Matlab code. It has been tested on computers running both Linux and Windows. It is important to note, that a job can be divided and allocated to *solvers* running under different operating systems and even under different versions of Matlab and Java Virtual Machines.

III. The installation and use of jPar

The *jPar* distribution package containing all the source files may be downloaded from the MatlabCentral Web page [9].

The installation of *jPar* is very simple:

1) To build *jPar* you need to have JDK (Java Development Kit) with `javac` and `jar` tools installed. Start command line (shell) interface and add JDK tools directory to your PATH environment variable if necessary (Windows) / change the proper shell configuration file (e.g.: `.profile`, `.zshrc`, `.cshrc`) (Unix/Linux).
2) Build the `jpar.jar` Java archive with command `"compile.bat"` (Windows) or `"sh compile.sh"` (Unix/Linux).
3) Copy the `.java.policy` file to home directory (in Windows use `"Documents and Settings\Username"` directory) or use `"install.bat"` (Windows) / `"sh ./install.sh"` (Unix/Linux) scripts on every node, where you want to run *jPar client* or *solver*.
4) Check whether the folder with Java binaries is in the system path writing `"java"` in the command window. If it is not, change the path variable, adding this directory (see p. 1).

The installation is performed only once. To use *jPar*, that is to start a session with it, you should:

1) Run one instance of *jPar server* using `"jpar_server.bat"` (Windows) / `"sh ./jpar_server.sh"` (Unix/Linux) on a node, where you want to run *jPar client*.
2) Start Matlab sessions and change the directory to the one which contains *jPar* files (if they were not started from this directory).
3) Start *solvers* from Matlab session in *jPar* directory using:

   ```
   >> jpar_solver(['<hostname>']);
   ```

   where `<hostname>` is the name of host where *jPar server* is running (default to `localhost`).
4) Start a distributed application by the following command:

   ```
   >> [<output>]=jpar_client(...
            '<name_of_the_function>',...
             <parameters>)
   ```

   where both input and output parameters are separated by commas.

After the work is done, the user should kill the *solvers*:

```
>> jpar_client('kill');
```

To see free *solvers* one may use the command:

```
>> jpar_client('hosts');
```

IV. Case study - parallel and distributed optimization

The tests have been performed on a network of Windows PCs with Intel Dual Core 3 GHz processors connected by 100 Mb/s network. All computers had common filesystem provided

by Linux server running *Samba*. The test job was to find the solution of a constrained separable optimization problem:

$$\min_{x\in X}\sum_{i=1}^{p} f_i(x_i) \tag{1}$$

subject to

$$\sum_{i=1}^{p} g_{ji}(x_i) \le M_j, \quad j=1,\ldots,m \tag{2}$$

$$x=(x_1,x_2,\ldots,x_p)\in X=X_1\times X_2\times\ldots X_p \tag{3}$$

$$X_i\subseteq\mathbb{R}^{n_i}, \quad n=\sum_{i=1}^{p} n_i \tag{4}$$

where all functions f_i are strictly convex, g_{ji} - convex, $i=1,\ldots,p; j=1,\ldots,m$.

Big optimization and optimal control problems solved in Matlab environment are often decomposed and parallelized [10]. To solve the above problem in a decomposed, parallel way the classical price method of hierarchical optimization was applied [11]. This method, which is often used to solve network optimization and control problems [12],[13], consists in the decomposition of the minimization of the Lagrangian $L(x,\lambda)$ while calculation of the dual function $L_D(\lambda)$ in the following way:

$$L_D(\lambda)=\min_{x\in X}\left[L(x,\lambda)=\sum_{i=1}^{p} f_i(x_i) + \sum_{j=1}^{m}\lambda_j\left(\sum_{i=1}^{p} g_{ji}(x_i)-M_j\right)\right]=$$

$$=\min_{\substack{x_i\in X_i,\\ i=1,\ldots,p}}\left[\sum_{i=1}^{p}\left(f_i(x_i)+\sum_{j=1}^{m}\lambda_j g_{ji}(x_i)\right)-\sum_{j=1}^{m}\lambda_j M_j\right]=$$

$$=\sum_{i=1}^{p}\min_{x_i\in X_i}\left(f_i(x_i)+\sum_{j=1}^{m}\lambda_j g_{ji}(x_i)\right)-\sum_{j=1}^{m}\lambda_j M_j \tag{5}$$

where λ_j, $j=1,\ldots,m$ are nonnegative Lagrange multipliers. In the natural way we obtain a hierarchical, two-level optimization scheme:

1) Local (slaves') level; the i-th local problem, $i=1,\ldots,p$:

$$\min_{x_i\in X_i}\left[L_i(x_i,\lambda)=f_i(x_i)+\sum_{j=1}^{m}\lambda_j g_{ji}(x_i)\right] \tag{6}$$

2) Coordination (master) level:

$$\max_{\lambda\ge 0}\left[L_D(\lambda)=\sum_{i=1}^{p} L_i(x_i(\lambda),\lambda)-\sum_{j=1}^{m}\lambda_j M_j\right] \tag{7}$$

where $x_i(\lambda)$ is the solution of the i-th local problem (6), $i=1,\ldots,p$.

The algorithm was implemented under Matlab with the hill climbing gradient method on the coordination level and the call of native Matlab `fmincon` solver on the local level. The tests were performed on the Powell20 problem [14]:

$$\min_{y\in R^n} 0.5(y_1^2+y_2^2+\ldots+y_n^2) \tag{8}$$

$$y_{k+1}-y_k\ge -0.5+(-1)^k\cdot k, \quad k=1,...,n-1 \tag{9}$$

$$y_1-y_n\ge n-0.5; \tag{10}$$

To transform this problem to the separable form (1)-(4) the vector x was divided into p parts of the dimension $n_1=n_2=\ldots=n_p=\frac{n}{p}$ (we assumed that $p|n$), what implied the corresponding division of the constraints: the p common ones were treated as global constraints and the Lagrange relaxation was applied to them (the remaining defined the subsequent sets X_i).

Denoting:

$$x_{i_j}=y_{(i-1)n_i+j},\ i=1,\ldots,p;\ j=1,...,n_i, \tag{11}$$

$$x_i=\left[x_{i_1},x_{i_2},\ldots,x_{i_{n_i}}\right]^T \tag{12}$$

$$\left[x_1^T,x_2^T,\ldots,x_p^T\right]^T=y \tag{13}$$

$$f_i(x_i)=0.5\sum_{k=1}^{n_i} x_{i_k}^2, \quad i=1,\ldots,p \tag{14}$$

$$X_i=\Big\{x_i\in\mathbb{R}^{n_i}|x_{i_l}-x_{i_{l+1}}-0.5+(-1)^{k(i,l)}\cdot k(i,l)\le 0, \quad l=1,\ldots,n_i-1\Big\} \tag{15}$$

where $k(i,l)=(i-1)\cdot n_i+l$; $c_j=-0.5+(-1)^{j\cdot n_j}\cdot(j\cdot n_j), j=1,\ldots,m$; $m=p$.

$$g_{ji}(x_i)\begin{cases} 0 & i\notin\{j,mod(j,p)+1\}\\ x_{i_{n_i}} & i=j\\ -x_{i_1} & i=mod(j,p)+1\end{cases} \tag{16}$$

for $i=1,\ldots,p; j=1,\ldots,m$.

$$M_j=-c_j=0.5-(-1)^{j\cdot n_j}\cdot(j\cdot n_j), \quad j=1,\ldots,m \tag{17}$$

we can transform our Powell20 problem (8)-(10) to the separable form (1)-(4).

What concerns the parallelization of the Matlab application code, the only necessary work was to replace the lines:

```
for i=1:p
    [xi(:,1,i),fi(1,1,i)]=pow20_pm_loct(...
        xloc(:,1,i),lambda,ni,p,options_k);
end
```

where `pow20_pm_loct` is the Matlab function solving the local problem (6), with the line:

```
[xi,fi]=jpar_client('pow20_pm_loct',...
                    xloc,lambda,ni,p,options_k);
```

TABLE I
TIMES OF CALCULATIONS WITH *Paralize* AND *jPar* (IN [S]).

n	p						
	1	2		4		6	
		Para-lize	jPar	Para-lize	jPar	Para-lize	jPar
108	1.68	17.7	16.1	12.7	15.8	23.9	21.7
216	14.68	39.5	40.4	33.1	31.5	40.4	41.0
432	58.82	366.8	331.6	100.6	99.5	89.9	84.5
648	2072	2033.2	1832.7	299.7	287.2	192.6	183.5
864	NT	2620.4	2390.5	810.1	786.8	381.3	377.7

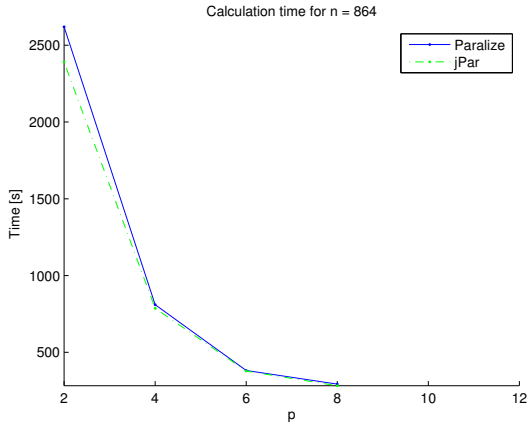


Fig. 2. The speed-up of the decomposition and parallelization for the biggest problem solved

The results are presented in Table I (NT means, that the calculations were not done for such (p, n) parameters) and in Fig. 2.
The superlinear speed-up in some tests was obtained due to a lucky choice of the starting points of optimization for some combinations of problem dimensions and the number of local problems.

Comparing to the original *Paralize*, *jPar* proved to be a bit faster, but more reliable and not "paralyzing" processors of the machines by empty loops.

V. CONCLUSIONS

We presented a very simple, portable package which may be used to parallelize Matlab scripts on multicore/multiprocessor computers with shared and local memory and in heterogenous clusters without a big effort. The only prerequisites are:

- fork-join structure of the application,
- access of all machines to the same file system.

The main advantages of *jPar* are:

- relatively small size,
- simplicity (the time spent on installation and learning it is very short),
- reliability (there are no errors caused by disk transmissions),
- heterogeneity (it was tested on x86 machines under both Linux and Windows),
- interoperability between various Matlab and Java versions,
- not blocking the machines between subsequent chunks of calculations and avoiding flooding of the local network with messages caused by active polling,
- openness, the free use with unlimited number of workers (*solvers*).

JPar is a little ($\sim$ 10%) faster than *Paralize*. Avoiding active polling it does not waste the energy and allows the cores to do other tasks.

Coarse-grained problems, that may be solved in the parallel way, are most suitable for the fast parallelization with *jPar*. The changes in code are very small and their best illustration is the line replacing the `for` loop block presented at the end of the Section IV.

A beta version of *j*Par was already used as the initial environment to perform parallel executions and the communication between master and slaves in MEIGO (MEtaheuristics for systems biology and bIoinformatics Global Optimization) [15] and CeSS (Cooperative enhanced Scatter Search) [16] packages, containing efficient solvers for hard global optimization problems arising in bioinformatics and computational systems biology, based on metaheuristics. This confirms that *jPar* is useful and easy to use not only for computer science specialists.

JPar is free and can be downloaded from the MatlabCentral page [9].

REFERENCES

[1] G. Sharma and J. Martin, "MATLAB®: A Language for Parallel Computing", International Journal of Parallel Programming, vol. 37, 2009, pp. 3–36, http://dx.doi.org/10.1007/s10766-008-0082-5
[2] J. Kepner, "Parallel Matlab for Multicore and Multinode Computers", SIAM Press, 2009, http://dx.doi.org/10.1137/1.9780898718126
[3] M. Buehren, "Multicore - Parallel processing on multiple cores", http://www.mathworks.com/matlabcentral/fileexchange/13775
[4] J. Kepner, "Parallel Programming with MatlabMPI", MIT Lincoln Laboratory, http://www.ll.mit.edu/mission/cybersec/softwaretools/matlabmpi/matlabmpi.html
[5] N. T. Bliss, J. Kepner, "pMatlab: Parallel Matlab Toolbox", MIT Lincoln Laboratory, http://www.ll.mit.edu/mission/cybersec/softwaretools/pmatlab/pmatlab.html
[6] J. Dillon, "sharedmatrix", http://www.mathworks.com/matlabcentral/fileexchange/28572
[7] T. Abrahamsson, "paralize (v2006a)", http://www.mathworks.com/matlabcentral/fileexchange/211
[8] Y.M. Altman, "Undocumented Secrets of MATLAB-Java Programming", CRC Press, 2011.
[9] A. Karbowski et al., "jPar - parallelizing Matlab calculations on multicores and in clusters without file communication", http://www.mathworks.com/matlabcentral/fileexchange/50797
[10] P. Drąg and K. Styczeń, "Parallel Simultaneous Approach for optimal control of DAE systems, in *Proc. Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 587–593.
[11] L.S. Lasdon, "Optimization theory for large systems", Macmillan, 1970; republished by Dover, 2002.
[12] S. Low, and D.E. Lapsley, "Optimization Flow Control, I: Basic Algorithm and Convergence", IEEE/ACM Transactions on Networking, vol. 7, 1999, pp. 861–874, http://dx.doi.org/10.1109/90.811451
[13] A. Karbowski, "Correction to Low and Lapsley's article "Optimization Flow Control, I: Basic Algorithm and Convergence", IEEE/ACM Transactions on Networking, vol. 11, 2003, pp. 338–339, http://dx.doi.org/10.1109/TNET.2003.810318

[14] M. J. D. Powell, “On the quadratic programming algorithm of Goldfarb and Idnani", Mathematical Programmimg Study, vol. 25, 1985, pp. 46–61, http://dx.doi.org/10.1007/BFb0121074

[15] J.A. Egea, D. Henriques, T. Cokelaer, A.F. Villaverde, A. MacNamara, D.-P. Danciu, J.R. Banga and J. Saez-Rodriguez, “MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics", BMC Bioinformatics, vol. 15, 2014, http://dx.doi.org/10.1186/1471-2105-15-136

[16] D.R. Penasa, P. González , J.A. Egea, J.R. Banga and R. Doallo, “Parallel Metaheuristics in Computational Biology: An Asynchronous Cooperative Enhanced Scatter Search Method", Procedia Computer Science, vol. 51, 2015, pp. 630–639, http://dx.doi.org/10.1016/j.procs.2015.05.331

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 97–104
DOI: 10.15439/2015F15
ACSIS, Vol. 6

Kaprekar's transformations. Part II – numerical results and intriguing corollaries

Edyta Hetmaniok, Mariusz Pleszczyński, Ireneusz Sobstyl, Roman Wituła
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: {edyta.hetmaniok,mariusz.pleszczynski,roman.witula}@polsl.pl

***Abstract*—This paper is a continuation of our previous paper [Part I, ibidem]. In this study we present many new results in the subject of minimal cycles (including the fixed points) of the so called Kaprekar's transformations. We formulate also some conjectures. Moreover, we discuss here all minimal cycles of the first 18 Kaprekar's transformations (and present but only of the first 15) with emphasis of the new, introduced by us, characteristics of this cycles.**

I. Introduction

In Part I of this elaboration (see [1]) we have introduced the definitions of the so called Kaprekar's transformations T_n:

$$\begin{aligned} T_n\colon \{0\}\cup\{\alpha\colon 10^{n-1}-1\le\alpha<10^n\} &\to \\ \to \{0\}\cup\{\alpha\colon 10^{n-1}-1\le\alpha<10^n\} &\\ T_n(\alpha) := \sum_{k=1}^{n}(a_k-a_{n-k+1})10^{k-1} &\\ = a_na_{n-1}\ldots a_1-a_1a_2\ldots a_n, & \end{aligned}$$

for every $\alpha, n\in\mathbb{N}$, $10^{n-1}-1\le\alpha<10^n$, where

$$0\le a_1\le a_2\le\ldots\le a_n\le 9,$$

denote all digits of decimal expansion of number α ordered in nondecreasing sequence and $T_n(0)=0$. We have also described the orbits of maps T_n for $n=3,4,\ldots,7$. Furthermore, in Part I many new concepts and characteristics of the minimal cycles of general transformations $F\colon X\to X$, where X is a finite set, have been proposed. All of them will be used in this part of our paper and applied for the Kaprekar's transformations T_n, $n\in\mathbb{N}$.

Moreover, in this part of our paper we intend to present firstly the collection of absolutely new facts discovered by observing the, numerically obtained, orbits of operators T_n for $n\le 18$. Next we will compile in tables the detailed descriptions of the minimal cycles of operators T_n for $n\le 15$ (that is, we will give many individual pieces of information concerning each of the investigated cycles). The other cases for $n=16-18$, because of the permissible length of the paper, are omitted here.

II. Facts basing on the numerical results

Let us present now several essential facts in the subject of Kaprekar's transformations which we have deduced by analyzing the numerically obtained minimal cycles of operators T_n for $n\le 18$. We will also formulate some conjectures concerning the cycles of Kaprekar's transformations.

Fact 1. *Numbers appearing in the orbits of transformations T_n correspond with the partitions of number $\lceil\frac{n}{2}\rceil\times 9$ into n digits, except the following $n=3k$-digit numbers being the Kaprekar's constants of order $3k$ with the sum of digits equal to $18k$:*

$$495,\ 549945,\ 554999445,\ \ldots,\ \underbrace{5\ldots5}_{(k-1)\text{ digits}}\ \underbrace{4\,9\ldots9}_{k\text{ digits}}\ \underbrace{4\ldots4}_{(k-1)\text{ digits}}\ 5.$$

The following theorem and the respective conclusions constitute the theoretical grounds of the described above properties of the orbits of transformations T_n.

Theorem 1.

a) *Let $a\in\mathbb{N}$ be a $2n$-digit number composed of digits*

$$0\le a_1\le a_2\le\ldots\le a_{2n}\le 9$$

and suppose that

$$a_{n-k-1}<a_{n-k}=a_{n-k+1}=\ldots=a_{n+l}<a_{n+l+1}$$

for some $k,l\in\mathbb{N}_0$.
If $k\ge l$, then the sum of digits of number $T_{2n}(a)$ is equal to $9\times(n+l)$. Otherwise, this sum is equal to $9\times(n+k)$.

b) *Let $a\in\mathbb{N}$ be a $(2n+1)$-digit number composed of digits*

$$0\le a_1\le a_2\le\ldots\le a_{2n+1}\le 9$$

and suppose that

$$a_{n-k}<a_{n-k+1}=a_{n-k+2}=\ldots=a_{n+l+1}<a_{n+l+2}$$

for some $k,l\in\mathbb{N}_0$.
If $k\ge l$, then the sum of digits of number $T_{2n+1}(a)$ is equal to $9\times(n+l+1)$, whereas if $k<l$, then the sum of digits of number $T_{2n+1}(a)$ is equal to $9\times(n+k+1)$.

Proof:
ad a) Let us notice that the following decimal expansions of $T_{2n}(a)$ can be obtained

$$T_{2n}(a)=\begin{cases}(a_{2n}-a_1)(a_{2n-1}-a_2)\ldots(a_{n+k+1}-a_{n-k}-1)\\ \times(9+a_{n+k}-a_{n-k})\ldots(9+a_2-a_{2n-1})\\ \times(10+a_1-a_{2n}),\ \text{if } l>k,\\ (a_{2n}-a_1)(a_{2n-1}-a_2)\ldots(a_{n+l+1}-a_{n-l}-1)\\ \times(9+a_{n+l}-a_{n-l})\ldots(9+a_2-a_{2n-1})\\ \times(10+a_1-a_{2n}),\ \text{if } k\ge l,\end{cases}$$

which implies the assertion.
ad b) The proof runs in similar way as in case of item a). ■

Corollary 1. *If $a \in \mathbb{N}$ is a n-digit number then the sum of digits of number $T_n(a)$ is not lower than the number $9 \times \lceil \frac{n}{2} \rceil$.*

Corollary 2. *If $a \in \mathbb{N}$ is a number possessing different digits in the decimal expansion then the sum of digits of number $T_n(a)$ is equal to $9 \times \lceil \frac{n}{2} \rceil$.*

Conjecture 1. *Sum of digits of the numbers belonging to the given orbit of operator T_n, where $n \in \mathbb{N}$, except the two-element orbit of operator T_5, is the same.*

Remark 1. *The lowest number n, for which there exist two different orbits (two different orbits possessing at least two elements) of operator T_n composed of the numbers with different sums of digits, is equal to 6 (is equal to $n = 16$, respectively).*

Remark 2. *Numbers belonging to the orbits of operator T_{2n+1} possess in their decimal expansion the middle digit equal to 9.*

Fact 2. *Let $a_1a_2 \dots a_n$ be an n-digit number belonging to some orbit of transformation T_n, $n \in \mathbb{N}$. Then the sequence, henceforward called as the digit type of element $a_1a_2 \dots a_n$ of the given cycle, defined in the following way*

$$a_1+a_n, a_2+a_{n-1}, a_3+a_{n-2}, \dots, \begin{cases} a_{\frac{n}{2}} + a_{\frac{n}{2}+1}, & \text{if } n \text{ is even} \\ a_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \end{cases}$$

is equal to

$$10, \underbrace{9, \dots, 9,}_{(k-1)\text{-times}} 8, 9 \tag{1}$$

if $n = 2k+1$, $k = 1, 2, \dots$, and

$$10, \underbrace{9, \dots, 9,}_{(k-2)\text{-times}} 8 \tag{2}$$

if $n = 2k$, $k = 2, 3, \dots$ In both cases the equality holds independently on number $a_1a_2 \dots a_n$, except the following numbers:

(i) *the Kaprekar's constants of order $n = 3k$:*

$$\underbrace{5 \dots 5}_{(k-1)\text{-times}} \; 4 \underbrace{9 \dots 9}_{k\text{-times}} \; \underbrace{4 \dots 4}_{(k-1)\text{-times}} \; 5,$$

for which the respective sequence of sums has the form

$$10, \underbrace{9, \dots, 9,}_{(k-2)\text{-times}} 8, \underbrace{18, \dots, 18,}_{\lfloor \frac{k}{2} \rfloor\text{-times}} \underbrace{9.}_{(\lceil \frac{k}{2} \rceil - \lfloor \frac{k}{2} \rfloor)\text{-times}}$$

Let us notice that if we correct the above sequence in the following way (we shift the units similarly as in the addition operation):

$$10, \underbrace{9, \dots, 9,}_{(k-2)\text{-times}} 8, \underbrace{18, 18, \dots, 18,}_{\lfloor \frac{k}{2} \rfloor\text{-times}} \underbrace{9.}_{(\lceil \frac{k}{2} \rceil - \lfloor \frac{k}{2} \rfloor)\text{-times}}$$

then we obtain the sequence

$$10, \underbrace{9, \dots, 9}_{(\lfloor \frac{3k}{2} \rfloor - 2)\text{-times}} 8, \underbrace{9,}_{(\lceil \frac{k}{2} \rceil - \lfloor \frac{k}{2} \rfloor)\text{-times}}$$

which is "compatible" either with (1), if k is odd, or with (2), if k is even.

(ii) *the numbers belonging to the single 2-element orbit $\{53955, 59994\}$ of operator T_5, where the respective sequences of sums are of the forms $10, 8, 9$ and $9, 18, 9$, but we get*

$$9, 18, 9 \longmapsto 10, 8, 9.$$

(iii) *the numbers belonging to the single 2-element orbit $\{8764421997755322, 8765431997654322\}$ of operator T_{16}, where both sequences of sums are of the form $10, 9, 9, 9, 9, 9, 8, 18$, but we obtain*

$$10,9,9,9,9,9,9,8, 18 \longmapsto 10,9,9,9,9,9,9,9,8,$$

which is compatible with (2).

Fact 3. *We have noticed that for every $n = 10, 12, ..., 18$ the operator T_n possesses the even number of 3-element cycles and, moreover, the difference between the numbers of 3-element cycles of T_n possessing the orbit types $(1, 3, 2)$ and $(1, 2, 3)$, respectively, is equal to 0 for $n = 10, 12$ and $2^{\frac{n-12}{2}}$ for $n = 14, 16, 18$. The orbit type of all 7-element cycles of T_n, $n \leq 18$, is the same and is equal to $(1, 5, 3, 4, 6, 7, 2)$.*

Fact 4 (Kaprekar's constants)**.** *We have observed that each Kaprekar's constant of order $n \leq 18$ generates the sequence of extensions of decimal expansions remaining the Kaprekar's constants (of the respectively higher order). For example, we have*

— $6 \underbrace{3 \dots 3}_{k\text{-times}} 17 \underbrace{6 \dots 6}_{k\text{-times}} 4$ *are the Kaprekar's constants of order $(2k+4)$ for every $k = 0, 1, 2, \dots$,*

Sketch of the proof: We have

$$7\underbrace{6 \dots 6}_{k+1}4\underbrace{3 \dots 3}_{k}1 - 1\underbrace{3 \dots 3}_{k}4\underbrace{6 \dots 6}_{k+1}7 = 6\underbrace{3 \dots 3}_{k}17\underbrace{6 \dots 6}_{k}4$$

■

— $\underbrace{9 \dots 9}_{k\text{-times}}750842\underbrace{0 \dots 0}_{(k-1)\text{-times}}1$ *are the Kaprekar's constants of order $(2k+6)$ for every $k = 1, 2, \dots$,*

— $975\underbrace{3 \dots 3}_{k\text{-times}}08\underbrace{6 \dots 6}_{k\text{-times}}421$ *are the Kaprekar's constants of order $(2k+8)$ for every $k = 0, 1, 2, \dots$,*

— $\underbrace{9 \dots 9}_{k\text{-times}}75308642\underbrace{0 \dots 0}_{(k-1)\text{-times}}1$ *are the Kaprekar's constants of order $(2k+8)$ for every $k = 1, 2, \dots$,*

— $864\underbrace{3 \dots 3}_{k\text{-times}}197\underbrace{6 \dots 6}_{k\text{-times}}532$ *are the Kaprekar's constants of order $(2k+9)$ for every $k = 0, 1, 2, \dots$*

Remark 3. *The Q-Kaprekar's transformations Q_n, defined in the last section of Part I, possess the same property as above for their fixed points. For example, the number*

$$\underbrace{5\dots5}_{k\text{-times}}4\,\underbrace{9\dots9}_{(k+1)\text{-times}}\,\underbrace{4\dots4}_{k\text{-times}}5$$

is the fixed point of transformation Q_{3k+3} for every $k = 1, 2, \dots$, the number

$$66\underbrace{3\dots3}_{k\text{-times}}08\underbrace{6\dots6}_{k\text{-times}}52$$

is the fixed point of Q_{2k+6} for every $k = 0, 1, 2, \dots$ and, at last, the number

$$\underbrace{9\dots9}_{(k+1)\text{-times}}750842\underbrace{0\dots0}_{k\text{-times}}1$$

is the fixed point of Q_{2k+8} for every $k = 1, 2, \dots$.

Fact 5. *We suppose that, similarly like in case of the Kaprekar's constants, all orbits of operators T_n with the odd number of elements possess their "extensions", that is they generate the infinite sequences of orbits of the Kaprekar's operators preserving the number of elements of the initial orbit. Whereas, despite of the insistent efforts we did not manage to get such extension (in the similar style as in case of the orbits presented below) for any orbit having the even number of elements.*

The Kaprekar's transformation $T_{2(k+4)}$, for $k = 0, 1, \dots, 5$, possesses $A140226(k)$ (equal to $\frac{1}{3}k(11+k^2)$ for $k \geq 1$) of 3-element minimal cycles ($A140226$ in notation of the Sloane's OEIS).

Furthermore, transformation $T_{2(k+4)}$, for each $k = 0, 1, \dots,$ possesses the following 3-element minimal cycle

$$\Big(643\underbrace{3\dots3}_{k\text{-times}}08\underbrace{6\dots6}_{k\text{-times}}654,\quad 83\underbrace{3\dots3}_{k\text{-times}}2087\underbrace{6\dots6}_{k\text{-times}}62,\quad 865\underbrace{3\dots3}_{k\text{-times}}26\underbrace{6\dots6}_{k\text{-times}}432\Big).$$

For $k = 0$ it is the single 3-element minimal cycle of the respective Kaprekar's transformation.

The other examples of 3-element minimal cycles of maps T_{6k+8}, T_{2k+10}, T_{2k+10}, are the following:

$$\Big(8\underbrace{7\dots7}_{k\text{-times}}\underbrace{3\dots3}_{2k\text{-times}}32087\underbrace{6\dots6}_{2k\text{-times}}62\underbrace{2\dots2}_{k\text{-times}},\quad 865\underbrace{5\dots5}_{k\text{-times}}\underbrace{3\dots3}_{2k\text{-times}}26\underbrace{6\dots6}_{2k\text{-times}}\underbrace{4\dots4}_{k\text{-times}}432,\quad 643\underbrace{3\dots3}_{2k\text{-times}}\underbrace{1\dots1}_{k\text{-times}}08\underbrace{8\dots8}_{k\text{-times}}\underbrace{6\dots6}_{2k\text{-times}}654\Big),$$

$$\Big(975\underbrace{3\dots3}_{k\text{-times}}1088\underbrace{6\dots6}_{k\text{-times}}421,\quad 9775\underbrace{3\dots3}_{k\text{-times}}08\underbrace{6\dots6}_{k\text{-times}}4221,\quad 9755\underbrace{3\dots3}_{k\text{-times}}08\underbrace{6\dots6}_{k\text{-times}}4421\Big),$$

$$\Big(975\underbrace{5\dots5}_{k\text{-times}}1088\underbrace{4\dots4}_{k\text{-times}}421,\quad 9775\underbrace{1\dots1}_{k\text{-times}}08\underbrace{8\dots8}_{k\text{-times}}4221,\quad 97\underbrace{7\dots7}_{k\text{-times}}550844\underbrace{2\dots2}_{k\text{-times}}21\Big),$$

respectively, for every $k = 0, 1, 2, \dots$

Every Kaprekar's transformation T_{2k+11}, for $k = 0, 1, 2, \dots$, possesses the following 5-element minimal cycle

$$\begin{aligned}&(864\,\underbrace{3\dots3}_{k-times}\,20987\,\underbrace{6\dots6}_{k-times}\,532,\\ &9664\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,5331,\\ &8843\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,6512,\\ &8764\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,5322,\\ &8654\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,5432).\end{aligned}$$

For $k = 0$ it is the single 5-element minimal cycle of the respective Kaprekar's transformation.

Next, the transformation T_{2k+13}, for every $k = 0, 1, 2, ...$, has also two following 5-element minimal cycles (all these cycles possess the same orbit type equal to $(1, 4, 5, 3, 2)$ and $(1, 4, 2, 5, 3)$, respectively):

$$\begin{aligned}&(8654\,\underbrace{3\dots3}_{k-times}\,20987\,\underbrace{6\dots6}_{k-times}\,5432,\\ &9664\,\underbrace{3\dots3}_{k-times}\,20987\,\underbrace{6\dots6}_{k-times}\,5331,\\ &98643\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,65311,\\ &88743\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,65212,\\ &87654\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,54322)\end{aligned}$$

and

$$\begin{aligned}&(8764\,\underbrace{3\dots3}_{k-times}\,20987\,\underbrace{6\dots6}_{k-times}\,5322,\\ &96654\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,54331,\\ &8843\,\underbrace{3\dots3}_{k-times}\,20987\,\underbrace{6\dots6}_{k-times}\,6512,\\ &97664\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,53321,\\ &88543\,\underbrace{3\dots3}_{k-times}\,197\,\underbrace{6\dots6}_{k-times}\,65412).\end{aligned}$$

For $k = 0$ three above 5-element minimal cycles are the only 5-element minimal cycles of T_{2k+13}.

The example of 7-element cycle of map T_{2k+6}, for every $k = 0, 1, 2, ...$, is the following (which possesses the orbit type

equal to $(1,5,3,4,6,7,2)$*):*

$$\Big(4\underbrace{3\ldots3}_{k\text{-times}}2087\underbrace{6\ldots6}_{k\text{-times}}6,\quad 85\underbrace{3\ldots3}_{k\text{-times}}17\underbrace{6\ldots6}_{k\text{-times}}42,$$
$$75\underbrace{3\ldots3}_{k\text{-times}}08\underbrace{6\ldots6}_{k\text{-times}}43,\quad 84\underbrace{3\ldots3}_{k\text{-times}}08\underbrace{6\ldots6}_{k\text{-times}}52,$$
$$86\underbrace{3\ldots3}_{k\text{-times}}08\underbrace{6\ldots6}_{k\text{-times}}32,\quad 86\underbrace{3\ldots3}_{k\text{-times}}26\underbrace{6\ldots6}_{k\text{-times}}32,$$
$$64\underbrace{3\ldots3}_{k\text{-times}}26\underbrace{6\ldots6}_{k\text{-times}}54\Big).$$

Indicated number 4, at the end of the last number in this cycle, appears only for $k \geq 1$.

For each $k \leq 6$ *this is the single 7-element minimal cycle of these Kaprekar's transformations.*

Fact 6. *The following statements hold for every* $n \leq 20$.

If T_n *possesses a cycle with the odd number of elements, then it possesses also a fixed point.*

Moreover, we note that there exists $n \leq 20$ *such that the operator* T_n *possesses only the nontrivial orbits with the even numbers of elements, for example we may consider* T_5*,* T_7.

Fact 7. *If* a *is an element belonging to the orbit of operator* T_n *composed of at least three numbers and* $a = \alpha_1\alpha_2\ldots\alpha_n$ *and* $T_n(a) = \beta_1\beta_2\ldots\beta_n$ *are the decimal representations of numbers* a *and* $T_n(a)$*, respectively, then* $\alpha_k - \beta_k = \beta_{n-k+1} - \alpha_{n-k+1}$ *for every* $k = 1,2\ldots,n$*. For example, for the cycles of operator* T_5 *(only two 4-element cycles are taken into account) we consider the following sequences of differences*

$$\beta_1 - \alpha_1,\ \beta_2 - \alpha_2,\ \ldots,\ \beta_5 - \alpha_5.$$

Thus, for the cycle

$$(62964 = a = T_5^4(a),\quad 71973 = T_5(a),$$
$$83952 = T_5^2(a),\quad 74943 = T_5^3(a))$$

we have

$$\underbrace{-1,-2,0,2,1}_{T_5^4(a)-T_5^3(a)};\quad \underbrace{1,-1,0,1,-1}_{T_5(a)-a};$$
$$\underbrace{1,2,0,-2,-1}_{T_5^2(a)-T_5(a)};\quad \underbrace{-1,1,0,-1,1}_{T_5^3(a)-T_5^2(a)},$$

whereas for the cycle

$$(61974, 82962, 75933, 63954)$$

we have

$$0,-2,0,2,0;\ 2,1,0,-1,-2;\ -1,3,0,-3,1;\ -1,-2,0,2,1.$$

III. Conclusions

Although one can find quite a lot of references concerning the subject of the discussed here Kaprekar's transformations (see the References in [1]), we have noticed yet several lacks in descriptions of the orbits of T_n transformations, even for $n \leq 10$. Aim of our work was to complete these lacks, in which we succeeded, and we did even more. Our achievements have been indicated and included in Section II. One should emphasize especially the theorems concerning the possibility of "expanding" the fixed points and cycles of a given Kaprekar's transformation T_n, $n \leq 18$, to the fixed points and cycles of infinitely many Kaprekar's transformations (which, by the way, gives the answer to a question whether there exist infinitely many $n \in \mathbb{N}$ such that T_n possesses a fixed point - similar fact concerns the possession of 3,5,7-element orbits). For our research we introduced several new concepts which, in the context of obtained numerical results, brought us to some theoretical results and conjectures. We derived some of our theorems and conjectures presented in Section II also for the generalizations of Kaprekar's transformations (obeying the Q-Kaprekar's transformation from [1]) which will be the subject of the created now next paper. We intend also to use the experience, gained by applying the numerical results in theory, in didactic work by showing to the students the possibilities of seemingly simple calculations. We will also use in this field the experiences of other authors (see [2], [3]).

Appendix

Description of tables presenting the cycles of Kaprekar's transformations T_n

The table is composed in the following way

- in the first row the value of index n of the Kaprekar's transformation T_n is given,
- the second row presents the amount of minimal cycles of the given length of the given transformation T_n as well as the information whether the given transformation preserves the strong Sharkovsky's order or the Sharkovsky's order (see definitions 1 and 2 in [1]),
- the third row shows how many n-digit numbers is transformed by the given Kaprekar's transformation T_n (after the finite number of steps) onto the respective minimal cycle of this transformation,
- in the successive rows the successive cycles from the third row (except the trivial one, that is the zero cycle) are associated with: the order types (it concerns only the cycles of length greater than 1, see the proper definition in [1]); the sum of digits of particular elements of the cycle, in case when these sums are identical, we include them only once; the digit types, and again, in case when they are identical, we include them only once; the longest increasing interval of the given cycle, the longest increasing subsequence of the given cycle, the longest decreasing interval of the given cycle and the longest decreasing subsequence of the given cycle.

References

[1] E. Hetmaniok, M. Pleszczyński, I. Sobstyl, R. Wituła, *Kaprekar's transformations I – theoretical discussion*, ibidem.

[2] U. Świerczynska-Kaczor and J. Wachowicz, "Student Response to Educational Games - An Empirical Study", *Proc. FedCSIS* (19th Conference on Knowledge Acquisition and Management), 2013, pp.1293-1299.

[3] N. S. Papaspyrou, S. Zachos, "Teaching programming through problem solving: The role of the programming language", *Proc. FedCSIS*, 2013, pp.1533-1536.

$n=5$				
	1 fixed point, 1 cycle of length 2, 2 cycles of length 4; strong Sharkovsky's order			
	3190 numbers $\to$ cycle: (53955,59994) 48480 numbers $\to$ cycle: (61974,82962,75933,63954) 48320 numbers $\to$ cycle: (62964,71973,83952,74943)			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1	$(1,2)$	$27, 36$	$(10,8,9)$, $(9,18,9)$	2, 1, 2, 1
β_2	$(1,4,3,2)$	27	$(10,8,9)$	2, 2, 3, 3
β_3	$(1,2,4,3)$	27	$(10,8,9)$	3, 3, 2, 2

$n=6$				
	3 fixed points, 1 cycle of length 7; Sharkovsky's order			
	1950 numbers $\to$ fixed point: 549945 62520 numbers $\to$ fixed point: 631764 935520 numbers $\to$ cycle: (420876,851742,750843,840852,860832,862632,64265)			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1		36	$(10,8,18)$	
β_2		27	$(10,9,8)$	
β_3	$(1,5,3,4,6,7,2)$	27	$(10,9,8)$	4, 5, 2, 3

$n=7$				
	1 fixed point, 1 cycle of length 8			
	9999990 numbers $\to$ cycle: (7509843,9529641,8719722,8649432,7519743,8429652,7619733,8439552)			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1	$(1,5,7,6,8,4,3,2)$	36	$(10,9,8,9)$	2, 4, 4, 5

$n=8$				
	3 fixed points, 1 cycle of length 3, 1 cycle of length 7			
	599536 numbers $\to$ fixed point: 63317664 2371040 numbers $\to$ fixed point: 97508421 48247316 numbers $\to$ cycle: (64308654,83208762,86526432) 48782098 numbers $\to$ cycle: (43208766,85317642,75308643,84308652,86308632,86326632,64326654)			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1, β_2		36	$(10,9,9,8)$	
β_3	$(1,2,3)$	36	$(10,9,9,8)$	3, 3, 1, 1
β_4	$(1,5,3,4,6,7,2)$	36	$(10,9,9,8)$	4, 5, 2 ,3

$n=9$				
	3 fixed points, 1 cycle of length 14; Sharkovsky's order			
	34440 numbers $\to$ fixed point: 554999445 51389136 numbers $\to$ fixed point: 864197532 948576414 numbers $\to$ cycle: $(753098643, 954197541, 883098612, 976494321, 874197522, 865296432, 763197633, 844296552, 762098733, 964395531, 863098632, 965296431, 873197622, 865395432)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1		54	$(10,9,8,18,9)$	
β_2		45	$(10,9,9,8,9)$	
β_3	$(1,11,10,14,9,6,3,4,2,12,5,13,8,7)$	45	$(10,9,9,8,9)$	2,5,4,6

$n = 10$				
	4 fixed points, 4 cycles of length 3, 1 cycle of length 7			
	4306680 numbers $\rightarrow$ fixed point: 6333176664 644450820 numbers $\rightarrow$ fixed point: 9753086421 41045760 numbers $\rightarrow$ fixed point: 9975084201 1291432626 numbers $\rightarrow$ cycle: $(6431088654, 8732087622, 8655264432)$ 3925269288 numbers $\rightarrow$ cycle: $(6433086654, 8332087662, 8653266432)$ 1058345520 numbers $\rightarrow$ cycle: $(6543086544, 8321088762, 8765264322)$ 558293820 numbers $\rightarrow$ cycle: $(9751088421, 9775084221, 9755084421)$ 2476855476 numbers $\rightarrow$ cycle: $(4332087666, 8533176642, 7533086643, 8433086652, 8633086632, 8633266632, 6433266654)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
$\beta_1 - \beta_3$		45	$(10, 9, 9, 9, 8)$	
β_4	$(1, 3, 2)$	45	$(10, 9, 9, 9, 8)$	2, 2, 2, 2
β_5, β_6	$(1, 2, 3)$	45	$(10, 9, 9, 9, 8)$	3, 3, 1, 1
β_7	$(1, 3, 2)$	45	$(10, 9, 9, 9, 8)$	2, 2, 2, 2
β_8	$(1, 5, 3, 4, 6, 7, 2)$	45	$(10, 9, 9, 9, 8)$	4, 5, 2 ,3

$n = 11$				
	2 fixed points, 1 cycle of length 5, 1 cycle of length 8			
	7444117296 numbers $\rightarrow$ fixed point: 86431976532 61796170458 numbers $\rightarrow$ cycle: $(86420987532, 96641975331, 88431976512, 87641975322, 86541975432)$ 30759712236 numbers $\rightarrow$ cycle: $(76320987633, 96442965531, 87320987622, 96653954331, 86330986632, 96532966431, 87331976622, 86542965432)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1		54	$(10, 9, 9, 9, 8, 9)$	
β_2	$(1, 5, 4, 3, 2)$	54	$(10, 9, 9, 9, 8, 9)$	2, 2, 4, 4
β_3	$(1, 6, 4, 8, 2, 7, 5, 3)$	54	$(10, 9, 9, 9, 8, 9)$	2, 3, 3, 4

$n = 12$				
	6 fixed points, 10 cycles of length 3, 1 cycle of length 7			
	697950 numbers $\rightarrow$ fixed point: 555499994445 57413664 numbers $\rightarrow$ fixed point: 633331766664 28903840680 numbers $\rightarrow$ fixed point: 975330866421 6771885120 numbers $\rightarrow$ fixed point: 997530864201 556839360 numbers $\rightarrow$ fixed point: 999750842001 23752825668 numbers $\rightarrow$ cycle: $(643110888654, 877320876222, 865552644432)$ 125925387258 numbers $\rightarrow$ cycle: $(643310886654, 873320876622, 865532664432)$ 250807302642 numbers $\rightarrow$ cycle: $(643330866654, 833320876662, 865332666432)$ 37978377360 numbers $\rightarrow$ cycle: $(654310886544, 873210887622, 876552644322)$ 124802255728 numbers $\rightarrow$ cycle: $(654330866544, 833210887662, 876532664322)$ 76745507520 numbers $\rightarrow$ cycle: $(655430865444, 832110888762, 877652643222)$ 14186684160 numbers $\rightarrow$ cycle: $(975110888421, 977750842221, 975550844421)$ 91728976482 numbers $\rightarrow$ cycle: $(975310886421, 977530864221, 975530864421)$ 35851244880 numbers $\rightarrow$ cycle: $(975510884421, 977510884221, 977550844221)$ 10397350260 numbers $\rightarrow$ cycle: $(997510884201, 997750842201, 997550844201)$ 171533411258 numbers $\rightarrow$ cycle: $(433320876666, 853331766642, 753330866643, 843330866652, 863330866632, 863332666632, 643332666654)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1		72	$(10, 9, 9, 8, 18, 18)$	
$\beta_2 - \beta_5$		54	$(10, 9, 9, 9, 9, 8)$	
β_6, β_7	$(1, 3, 2)$	54	$(10, 9, 9, 9, 9, 8)$	2, 2, 2, 2
$\beta_8 - \beta_{11}$	$(1, 2, 3)$	54	$(10, 9, 9, 9, 9, 8)$	3, 3, 1, 1
β_{12}, β_{13}	$(1, 3, 2)$	54	$(10, 9, 9, 9, 9, 8)$	2, 2, 2, 2
β_{14}	$(1, 2, 3)$	54	$(10, 9, 9, 9, 9, 8)$	3, 3, 1, 1
β_{15}	$(1, 3, 2)$	54	$(10, 9, 9, 9, 9, 8)$	2, 2, 2, 2
β_{16}	$(1, 5, 3, 4, 6, 7, 2)$	54	$(10, 9, 9, 9, 9, 8)$	4, 5, 2, 3

$n = 13$				
	2 fixed points, 1 cycle of length 2, 3 cycles of length 5; Sharkovsky's order			
	127766869230 numbers $\to$ fixed point: 8643319766532 729214292326 numbers $\to$ cycle: $(8733209876622, 9665429654331)$ 5169476073242 numbers $\to$ cycle: $(8643209876532, 9664319765331, 8843319766512, 8764319765322, 8654319765432)$ 1373689940636 numbers $\to$ cycle: $(8654209875432, 9664209875331, 9864319765311, 8874319765212, 8765419754322)$ 2599852824556 numbers $\to$ cycle: $(8764209875322, 9665419754331, 8843209876512, 9766419753321, 8854319765412)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1		63	$(10, 9, 9, 9, 9, 8, 9)$	
β_2	$(1, 2)$	63	$(10, 9, 9, 9, 9, 8, 9)$	2, 2, 1, 1
β_3	$(1, 5, 4, 3, 2)$	63	$(10, 9, 9, 9, 9, 8, 9)$	2, 2, 4, 4
β_4	$(1, 4, 5, 3, 2)$	63	$(10, 9, 9, 9, 9, 8, 9)$	3, 3, 3, 3
β_5	$(1, 4, 2, 5, 3)$	63	$(10, 9, 9, 9, 9, 8, 9)$	2, 3, 2, 2

$n = 14$				
	7 fixed points, 20 cycles of length 3, 1 cycle of length 7			
	825128304 numbers $\to$ fixed p.: 63333317666664; 1640938809510 numbers $\to$ fixed p.: 97533308666421 1955480289854 numbers $\to$ fixed p.: 97755108844221; 516356961120 numbers $\to$ fixed p.: 99753308664201 126071225280 numbers $\to$ fixed p.: 99975308642001; 6034588560 numbers $\to$ fixed p.: 99997508420001 616791947798 numbers $\to$ cycle: $(64311108888654, 87773208762222, 86555526444432)$ 2245517211436 numbers $\to$ cycle: $(64331108886654, 87733208766222, 86555326644432)$ 12115951630042 numbers $\to$ cycle: $(64333108866654, 87333208766622, 86553326664432)$ 20900682225326 numbers $\to$ cycle: $(64333308666654, 83333208766662, 86533326666432)$ 1233797593392 numbers $\to$ cycle: $(65431108886544, 87732108876222, 87655526444322)$ 4978650152970 numbers $\to$ cycle: $(65433108866544, 87332108876622, 87655326644322)$ 8893048070816 numbers $\to$ cycle: $(65433308666544, 83332108876662, 87653326664322)$ 1917234715396 numbers $\to$ cycle: $(65543108865444, 87321108887622, 87765526443222)$ 4466367674132 numbers $\to$ cycle: $(65543308665444, 83321108887662, 87765326643222)$ 1355384297358 numbers $\to$ cycle: $(65554308654444, 83211108888762, 87776526432222)$ 360886383858 numbers $\to$ cycle: $(97511108888421, 97777508422221, 97555508444421)$ 2896580093862 numbers $\to$ cycle: $(97531108886421, 97775308642221, 97555308644421)$ 5677743145438 numbers $\to$ cycle: $(97533108866421, 97753308664221, 97553308664421)$ 2626503498710 numbers $\to$ cycle: $(97551108884421, 97775108842221, 97755508444221)$ 6197474439338 numbers $\to$ cycle: $(97553108864421, 97753108864221, 97755308644221)$ 1366108585842 numbers $\to$ cycle: $(97555108844421, 97751108884221, 97775508442221)$ 420203255472 numbers $\to$ cycle: $(99751108884201, 99777508422201, 99755508444201)$ 2316236914992 numbers $\to$ cycle: $(99753108864201, 99775308642201, 99755308644201)$ 829988923764 numbers $\to$ cycle: $(99755108844201, 99775108842201, 99775508442201)$ 181449067800 numbers $\to$ cycle: $(99975108842001, 99977508422001, 99975508442001)$ 14157693169620 numbers $\to$ cycle: $(43333208766666, 85333317666642, 75333308666643, 84333308666652, 86333308666632, 86333326666632, 64333326666654)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
$\beta_1 - \beta_6$		63	$(10, 9, 9, 9, 9, 9, 8)$	
$\beta_7 - \beta_9$	$(1, 3, 2)$	63	$(10, 9, 9, 9, 9, 9, 8)$	2, 2, 2, 2
β_{10}, β_{25}	$(1, 2, 3)$	63	$(10, 9, 9, 9, 9, 9, 8)$	3, 3, 1, 1
β_{11}, β_{26}	$(1, 3, 2)$	63	$(10, 9, 9, 9, 9, 9, 8)$	2, 2, 2, 2
$\beta_{12} - \beta_{16}$	$(1, 2, 3)$	63	$(10, 9, 9, 9, 9, 9, 8)$	3, 3, 1, 1
$\beta_{17} - \beta_{20}$	$(1, 3, 2)$	63	$(10, 9, 9, 9, 9, 9, 8)$	2, 2, 2, 2
β_{21}, β_{22}	$(1, 2, 3)$	63	$(10, 9, 9, 9, 9, 9, 8)$	3, 3, 1, 1
β_{23}, β_{24}	$(1, 3, 2)$	63	$(10, 9, 9, 9, 9, 9, 8)$	2, 2, 2, 2
β_{27}	$(1, 5, 3, 4, 6, 7, 2)$	63	$(10, 9, 9, 9, 9, 9, 8)$	4, 5, 2, 3

$n = 15$				
	3 fixed points, 1 cycle of length 2, 5 cycles of length 5			
	15165150 numbers $\rightarrow$ fixed p.: 555549999944445; 3577552068090 numbers $\rightarrow$ fixed p.: 864333197666532 12790914986700 numbers $\rightarrow$ cycle: $(873332098766622, 966543296654331)$ 91463039030240 numbers $\rightarrow$ cycle: $(864332098766532, 966433197665331, 884333197666512, 876433197665322, 865433197665432)$ 234193123825336 numbers $\rightarrow$ cycle: $(865432098765432, 966432098765331, 986433197665311, 887433197665212, 876543197654322)$ 270342559594928 numbers $\rightarrow$ cycle: $(876432098765322, 966543197654331, 884332098766512, 976643197653321, 885433197665412)$ 146805971092664 numbers $\rightarrow$ cycle: $(876542098754322, 966542098754331, 986432098765311, 987643197653211, 887543197654212)$ 240826824236882 numbers $\rightarrow$ cycle: $(885432098765412, 976642098753321, 986543197654311, 887432098765212, 976654197543321)$			
successive cycles	order type	sum of digits	digit type	longest incr. interval, subseq., longest decr. interval, subseq.
β_1		90	$(10, 9, 9, 9, 8, 18, 18, 9)$	
β_2		72	$(10, 9, 9, 9, 9, 9, 8, 9)$	
β_3	$(1, 2)$	72	$(10, 9, 9, 9, 9, 9, 8, 9)$	2, 2, 1, 1
β_4	$(1, 5, 4, 3, 2)$	72	$(10, 9, 9, 9, 9, 9, 8, 9)$	2, 2, 4, 4
β_5	$(1, 4, 5, 3, 2)$	72	$(10, 9, 9, 9, 9, 9, 8, 9)$	3, 3, 3, 3
β_6	$(1, 4, 2, 5, 3)$	72	$(10, 9, 9, 9, 9, 9, 8, 9)$	2, 3, 2, 2
β_7	$(1, 3, 4, 5, 2)$	72	$(10, 9, 9, 9, 9, 9, 8, 9)$	4, 4, 2, 2
β_8	$(1, 3, 5, 2, 4)$	72	$(10, 9, 9, 9, 9, 9, 8, 9)$	3, 3, 2, 2

2nd International Workshop on Cyber-Physical Systems

PROLIFERATION of computers in everyday life requires cautious investigation of approaches related to the specification, design, implementation, testing, and use of modern computer systems interfacing with real world and controlling their environment. Cyber-Physical Systems (CPS) are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. Cyber-physical systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The event is a continuation and extension of 2006-2010 Real-Time Software FedCSIS workshops and 2013 IWCPS. The objective of the workshop is to assemble and develop a community with main interest in cyber-physical systems.

Topics

Due to an extensive scope of the topics, the workshop will accept papers in the following areas:

- Control Systems
  - embedded/networked/intelligent
  - wireless sensing/actuation
  - adaptive/predictive
- Scalability/Complexity
  - modularity
  - design methodology
  - legacy systems
  - tools
- Interoperability
  - concurrency
  - models of computation
  - networking
  - heterogeneity
- Validation and Verification
  - assurance
  - certification
  - simulation
- Cyber-security
  - intrusion detection
  - resilience
  - privacy
  - attack vectors
- Applications of CPS
  - robotics
  - transportation
  - military
  - medical
  - consumer
  - manufacturing
  - power systems
- CPS Education
  - curriculum development
  - web-based laboratories
  - academic courses
  - pedagogy issues

Event Chairs

Grega, Wojciech, AGH University of Science and Technology, Poland

Kornecki, Andrew J., Embry Riddle Aeronautical University, United States

Nigro, Libero, Università della Calabria, Italy

Szmuc, Tomasz, AGH University of Science and Technology, Poland

Zalewski, Janusz, Florida Gulf Coast University, United States

Program Committee

Angiulli, Fabrizio, University of Calabria

Babiceanu, Radu, ERAU

Cicirelli, Franco, Università della Calabria, Italy

Crespo, Alfons, Universitat Politecnica de Valencia, Spain

Golatowski, Frank, University of Rostock, Germany

Gomes, Luis, Universidade Nova de Lisboa, Portugal

Halang, Wolfgang A., Fernuniversitaet, Germany

Letia, Tiberiu, Technical University of Cluj-Napoca, Romania

Malec, Jacek, Lund University, Sweden

Marwedel, Peter, Technische Universität Dortmund, Germany

Motus, Leo, Tallinn University of Technology, Estonia

Nadjm-Tehrani, Simin, Linköping University, Sweden

Nigro, Libero, Universite della Calabria

Rysavy, Ondrej, Brno University of Technology, Czech Republic

Sanden, Bo, Colorado Technical University, United States

Schagaev, Igor, London Metropolitan University, United Kingdom

Seker, Remzi, Embry Riddle Aeronautical University, United States

Sveda, Miroslav, Brno University of Technology, Czech Republic

Trybus, Leszek, Rzeszow University of Technology, Poland

Vardanega, Tullio, University of Padova, Italy

Villa, Tiziano, Università di Verona, Italy

Zoebel, Dieter, University Koblenz-Landau, Germany

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 107–114
DOI: 10.15439/2015F414
ACSIS, Vol. 6

Modeling Resiliency and Its Essential Components for Cyberphysical Systems

Janusz Zalewski
Software Engineering Dept.
Florida Gulf Coast Univ.
Ft. Myers, FL 33965
USA
zalewski@fgcu.edu

Steven Drager
William McKeever
Air Force Research Lab
Rome, NY 13441, USA
Steven.Drager@us.af.mil,
William.McKeever.1@us.af.mil

Andrew J. Kornecki
ECSSE Department
Embry-Riddle Aero. Univ.
Daytona Beach, FL 32114
USA
kornecka@erau.edu

Bogdan Czejdo
Dept. of Math. & CS
Fayetteville State Univ.
Fayetteville, NC 28301
USA
bczejdo@uncfsu.edu

***Abstract*—This paper presents an initial approach related to modeling resiliency for cyberphysical systems. It discusses the concept and definitions of resiliency and outlines the process of building a model of resiliency. Through analogies with feedback control and fault tolerance, the Design for Resilience is addressed, where the design of the controller component of a cyberphysical system needs to account for potential safety hazards and security threats, with awareness of its internal faults and vulnerabilities. This model is validated against other approaches to modeling resilience described in the literature, followed by a discussion of the resilience metrics. The paper concludes with presenting the strategy of modeling resiliency, based on the assumption that one cannot guarantee absolute protection against attacks, or failures, but can aim at providing successful recovery after disruptions. With safety and security as essential resiliency components, an extended model is proposed involving an attacker, suggesting appropriate performance metric reflecting the distance between the normal state and the degraded state. A model-based environment Möbius, from the University of Illinois, is considered in helping to evaluate resiliency under various operational scenarios.**

I. Introduction

ALTHOUGH resiliency is essentially a concept adopted in medicine and health science [1] and relates to patient's resistance in response to disease, in common sense, resiliency (or resilience) is often associated with natural or ecological systems demonstrating tolerance to, and respective recovery from, disasters, such as earthquakes, floods, hurricanes, etc. [2]. The concept has been also extended to human-made environments, such as supply chains [3], transportation networks [4], military operations [5], etc., which are called resilient if they can tolerate some major failures or disruptions and smoothly return to normal operational capability. Recent books in systems engineering take note of additional aspects of resiliency, including redundancy [6], adaptability [7], and safety as the ability to succeed under varying conditions [8].

This project has been funded in part by the 2014 Visiting Faculty Research Program at the Air Force Rome Labs. Case Number 88ABW-2015-3306 – 26 June 2015. Distribution unlimited.

In computing, and in cyberphysical systems in particular, the term resilience has been adopted to describe the computer system's ability to restore its original functionality after a loss [9]-[11]. In a contemporary world, it concerns primarily computer networks and cybersecurity, applied in various types of systems, from critical infrastructure [12] to space systems [13], and more.

Resiliency landscape up to early 2011 has been covered in MITRE report [14], which listed approximately 320 articles, divided in eight categories, one of them particularly relevant to modeling, resiliency metrics. In current work, a number of more recent papers (dated 2011 and later) were analyzed, with respect to models of resiliency.

The particular objective of this work is to address resiliency assessment of cyberphysical systems in response to multiple external disturbances and internal parameter fluctuations, as follows:

- Identify critical components of resiliency, beyond security, such as reliability, safety, etc.
- Develop a process for resiliency assessment in cyberphysical systems.
- Apply it to the control-theoretic model of a resilient architecture to assess its resiliency.

The rest of the paper is structured as follows. Section II discusses in detail the concept of resiliency, Section III describes the adopted model of resiliency, followed by its expansion in Section IV and a discussion of metrics and measures in Section V. Section VI presents the modeling strategy, and Section VII constitutes the conclusion.

II. The Concept of Resiliency

To set the stage for serious research on resiliency, some fundamental issues of understanding the concept must be resolved. For example, some authors [12] look at the assessment of resiliency (calling it resilience) from two perspectives: design methods and system operation:

Cyber resilience design methods consider primarily how system architecture and activities enhance the resilience of the system to cyber threats. The second

category, operational resilience assessment methods […] consider physical threats and accidents, in addition to cyber threats.

The architectural view is advocated in [15], stating that "Architectural resiliency is the ability of an architecture – for an enterprise, a mission / business segment, a system-of-systems, a family of systems, or an individual system or component – to enable missions (including cyber defense missions) to anticipate, withstand, recover from, and evolve to address more effectively, cyberdomain attacks."

On the other hand, operational resilience is discussed in detail in [10], as opposed to enterprise resilience, which in fact is consistent with the architectural perspective. Operational resilience, adopting definition from [16], is understood as "the organization's ability to adapt to risk that affects its core operational capacities. […] A subset of enterprise resilience, operational resilience, focuses on the organization's ability to manage operational risk, whereas enterprise resilience encompasses additional areas of risk such as business risk and credit risk".

In a different perspective, Madni and Jackson [17] describe resilience as the ability to bounce back after a shock or disturbance, and consider it as a "multi-faceted capability of a complex system that encompasses avoiding, absorbing, adapting to, and recovering from disruptions."

To summarize various understandings of resilience, one can see that some views consider resilience as a state of a system, and some others see it as a system property, calling it ability. These two different, although overlapping, ways of studying resilience are adopted in this work. One notion of resiliency relating to the concept of system state encompasses the system architecture view and operational view, and answers the question:

How to build or operate a computing system to make it resilient?

The second notion of resiliency relates to it as an ability, or system property (attribute), and answers a different question:

To what extent (or to what degree), an existing computing system is resilient?

This dual understanding of a concept of resiliency, one based on studying system state (its architecture and/or operation) and the other based on studying a system attribute or property is adopted in this work, with focus on studying and modeling resiliency as a property.

Such dualism in understanding a system related concept is not that uncommon, as it may look at the first glance, and has further consequences. Since there are essentially two notions, two different definitions of these concepts are needed, and possibly two different terms to denote it:

- After [18], resilience is defined as: the ability to maintain acceptable levels of operation in the presence of abnormal conditions.
- Following this definition, we define resiliency as: the extent to which a computing system is able to maintain acceptable levels of operation in the presence of abnormal conditions.

These definitions fit into multiple others encountered in the literature. For example, Meyer defines resilience as the persistence of performability when facing changes [19]. Bishop claims [20] that "a resilient system is effectively a survivable system that is capable of restoring not only its performance level back to desired levels, but also the capacity of the system itself to recover, maintaining its ability to sustain future attacks or failures." Others define resilience or resiliency as the ability of restoring original operational capabilities (functionality) after a loss [12]. An overall consensus seems to be that resilience characterizes system's ability to conduct recovery from serious disruption.

As a final remark, regarding the subtle distinction between the notions of resilience and resiliency, one has to mention that the split into two separate notions, one based on system state and the other based on system property, is not specific to resiliency. A similar situation exists, although is rarely articulated, with the concept of security, where in addition to security understood as related to system state, there is a concept of security as a system property. The same situation exists with the two concepts of safety. In these cases, it would be proper to coin a different term for one of those close meanings, and talk about *security* and *secureness* and, correspondingly, about *safety* and *safeness*.

III. Building a Model of Resiliency

A. General Considerations

Building a model of resiliency to explore it is not a new topic. Older papers, referred to in the MITRE study [14], seem to discuss it in general terms, at the conceptual level, without even using the term model. There has been also a variety of papers published over the years, how to approach studying resilience (resiliency) and building models from the point of view called resilience engineering, for example [5], [17], [21]. All these models, however, are primarily conceptual and do not facilitate quantitative, or even qualitative, analysis of resiliency. The major shortcoming of all of such attempts and their corresponding models is the lack of mathematical underpinning.

One point that everyone agrees upon, because it is inherent in essentially all definitions of resilience or resiliency, is the illustration of divergence from desired operational conditions due to a sudden disruption, and successful recovery to the desired state, as shown in Figure 1, adopted from [20]. The model is expressed in terms of Quality of Service (QoS) varying over time, and represents a dip in performance, understood as diverging from specified operational conditions, which is caused by a sudden disruption at time A. Value of (B-A) represents the time taken for the system to return to its equilibrium state E. Value of (E-C) represents the maximum disturbance for system marked in blue. Another possible response is shown

for system marked in green. Point F represents a QoS below which the system's mission is compromised [20].

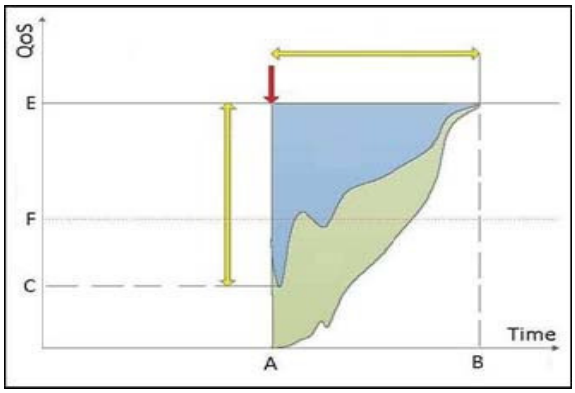


Fig. 1 Illustration of a concept of resilience to a sudden disruption

This model, being relatively widely adopted, is very illustrative for our purposes, because it splits the concept of resilience into its constituting components:

- the system's equilibrium state E
- disturbance or disruption point A
- acceptable service degradation level F
- maximum divergence from the equilibrium, E-C, and
- length of time interval to return to equilibrium, B-A.

The question is now, how to determine these points and intervals, and – once they are determined – how to develop behavioral policies or operational principles for the system to possibly anticipate the potential disruption and to respond to sudden disruptions and recover from degradation of state to fully operational conditions. The rest of this section outlines the process of building such a model.

B. Analogies with Feedback Control and Fault Tolerance

Feedback Control Analogy. Looking at the illustration in Figure 1, one can immediately find a behavioral analogy with a typical feedback control system, which is shown in Figure 2. For such system, any disruption caused by disturbances results in changes of the Measured Value, which cause its deviation from the Setpoint, represented as ε. The Controller then is responsible for following the Control Law (an algorithm, which determines respective action) and sending an appropriate Control Signal to the Controlled Object to return it to the equilibrium state, as indicated by the Measured Value. Thus, the analogy with the concept of resilience illustrated in Figure 1 can be described as follows:

- the Setpoint (desired value) in Figure 2 corresponds to the system's equilibrium E, in Figure 1
- disturbances in Figure 2 correspond to the disruption at point A in Figure 1
- the deviation from the Setpoint, ε, in Figure 2, corresponds to divergence from the equilibrium, E-C, in Figure 1
- the time constant, τ, for the control system in Figure 2, corresponds to time interval to return to equilibrium, B-A, and
- parameters such as overshoot for the control system in Figure 2 may be viewed as corresponding to acceptable service degradation level F in Figure 2.

This analogy is very instructive not only as a simple illustration of concepts. Its primary result is the conceptual formulation of the Design for Resilience problem.

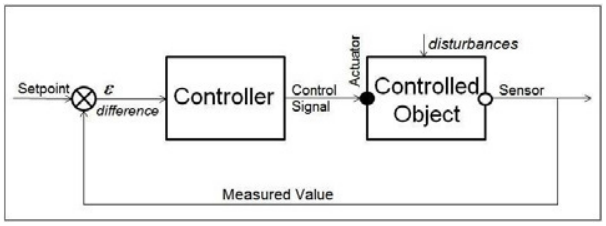


Fig. 2 Illustration of a control system influenced by disturbances

A typical control problem for the system shown in Figure 2 may be articulated as follows. For a given Controlled Object and Disturbances, design a Controller to generate Control Signal that minimizes certain characteristics of the Controlled Object expressed in terms of a Criterion (performance index) usually formulated in terms of the difference, **ε**, between the Setpoint and the Measured Value.

Obviously, strict formulation as mathematical description is needed for both the Controlled Object and Disturbances, as well as for the Criterion used as an indicator of the performance of the Controller. Then, a Control Law can be derived using, e.g., linear feedback control theory [22].

With this in mind, the problem of Design for Resilience (Feedback Control Analogy) can be formulated as follows.

Given (1) the description of the System whose resilience is of concern (analog of the Controlled Object), and (2) the characteristic of the expected Disruptions, develop a Strategy (an analog of a Control Law running on the Controller) to meet a certain Criterion (performance metric) expressed in terms of the distance from the desired state of the System.

There are more analogies between feedback control systems and resilient systems, stemming mostly from the fact that feedback control is very naturally illustrating resilience. For example, to understand effects of sudden disruptions on control systems and draw further analogies with resilient systems, one can talk about step response and impulse response functions [20], tolerating single or multiple upsets, and so on.

Fault Tolerance Analogy. Feedback control deals mostly with response to external disturbances, which are assumed to negatively affect the Controlled Object (Figure 1), be random and well characterized mathematically (for example, described by a Gaussian noise). However, all modern control systems are nowadays implemented digitally, and are

significantly expanded dealing with a User (Operator), are connected to the Network, as well as to a Database, which may be viewed as a logical extension of a single Setpoint data value. This is illustrated in Figure 3. With this complexity of controller interactions, when designing a Controller one has to take into account Controller's internal state, which may be a cause of significant disruptions to the Controlled Object, when a Controller fails. This is the subject of fault tolerance.

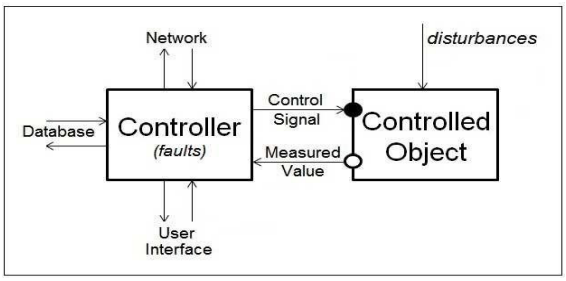


Fig. 3 Illustration of interactions in a modern control system

Fault Tolerance (FT) is a well-developed research domain [23], which has produced numerous methods, techniques and tools to deal with faults and failures. Some of the methods include: graceful degradation, diversity, redundancy, N-version programming, fail safety, and others [20]. In particular, the techniques related to FT are widely applied in dealing with faults to improve safety of cyberphysical systems: Fault Tree Analysis (FTA), Event Tree Analysis (ETA), Failure Mode and Effect Analysis (FMEA), as well as other techniques, such as Markov chains, Petri nets, Hazard and Operability Analysis (HAZOP), etc.

The subject of fault tolerance has been discussed in association with resilience, beginning as early as in 1990 [24]. One specific approach worth mentioning relates it to safety. The paper [17] states that to achieve resilience "The primary implication of external disruptions is that systems need to be built with adequate safety margins to account for uncertainty." Technically, in safety engineering, external disruptions are representing hazards and in the model from Figure 3 can be viewed as affecting the Controlled Object, as specific disturbances. Formally, a hazard is an intrinsic property or condition that has the potential to cause harm or damage [25]. To assure resilience, the Controller has to be designed to deal with safety hazards, but they are not always easy to capture and are especially difficult to account for in case of hardware or software faults.

Assuming that a fault in the Controller hardware or software, when activated, may cause a failure that will negatively affect the behavior of the Controlled Object, one can reformulate the Design for Resilience (Fault Tolerance Analogy) problem as follows.

Given (1) the description of the System whose resilience is of concern, (2) the characteristic of the expected external Disruptions, including Hazards, and (3) the characteristic of internal Faults, develop a Strategy (an analog of a Control Law running on the Controller) to meet a certain Criterion (performance metric) expressed in terms of the distance from the desired state of the System.

C. Including Security

When dealing with resilience one has to keep in mind that such discussions always involve cybersecurity [25]-[27], which is nowadays considered a primary factor in studying resilience. Nevertheless, any discussion involving security issues and its relationship to resilience is usually self-contained and almost never involves one other important constituting factor of resilience, which is safety.

One has to remember, however, that security and safety are two sides of the same coin, mutually complementary aspects of resilience. According to the International Electrotechnical Commission (IEC) [28], safety is defined as "freedom from unacceptable risk to the outside from the functional and physical units considered" whereas security is defined as "freedom from unacceptable risk to the physical units considered from the outside." Translating this into the language used in the current report:

- Safety is concerned when a Controller failure leads to severe consequences (high risk) to the environment (including Controlled Object);
- Security is concerned when a Controller failure to protect assets (a breach) leads to severe consequences (high risk) to the Controller itself (and potentially to the Controlled Object).

There are numerous definitions of security as a system property, but the one that is the most valuable should include the C+I+A (Confidentiality, Integrity and Availability) factors. In this view, the definition adopted from [29] reads as follows:

Security - the extent to which information and data are protected so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them.

A key element in this definition is "unauthorized access." From the perspective of protecting the system, this unauthorized access is called a threat. A corresponding definition taken from [28] reads as follows:

Threat - a state of the system or system environment which can lead to adverse effect in one or more given risk dimensions.

Assuming that a threat comes from the environment, as in the definition above, one can reflect it in the adjusted diagram of the control system used in the model of resilience (Figure 4). The new diagram shows that multiple Controller interfaces, the one to the Controlled Object, those to the User (Operator), the Network, and the Database, are all subject to security threats, thus forming the attack surface.

More importantly, to take the analogy further, just like control theory assumes that the Controlled Object is subject to Disturbances, security theory, if one is developed for this model, or resilience engineering, could assume that known or unknown Threats play the role of Disturbances to the Controller. Threats can only be effective if they exploit some weaknesses of the Controller called vulnerabilities. In this model, vulnerabilities affecting the controller are endangering the system assets that can be exploited by one or more threats. The formal definition [30] reads as follows:

Vulnerability – a weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source.

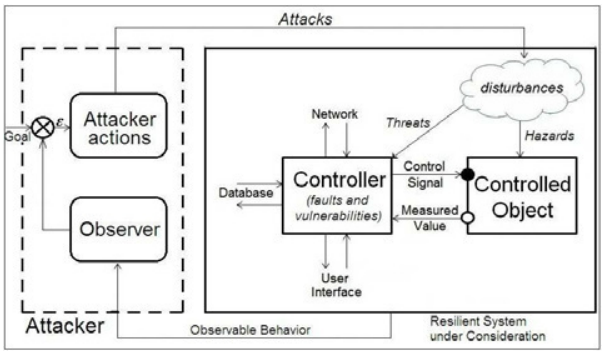


Fig. 4 Modern control system: disturbances and attacks

Pairing this understanding of security related concepts of Threats and Vulnerabilities with safety related concepts of Hazards and Faults, one arrives to the aggregated model suitable for resilience modeling, as shown in Figure 4. Assuming further that existing vulnerabilities in the Controller hardware or software, when exploited, may cause a security breach negatively affecting the behavior of the Controller, one can formulate the Design for Resilience Considering Security problem as follows:

Given (1) the description of the System whose resilience is of concern, (2) the characteristic of the expected external Disruptions, including Hazards and Threats, (3) the characteristics of internal Faults and Vulnerabilities, develop a Strategy (an analog of a Control Law running on the Controller) to meet a certain Criterion (performance metric) expressed in terms of the distance from the desired state of the System.

IV. Verification and Expansion of the Model

The Verification Problem can be formulated as follows:

Given (1) the description of the System whose resilience is of concern, (2) the characteristic of the expected external Disruptions, including Hazards and Threats, (3) the characteristics of internal Faults and Vulnerabilities, develop a Strategy (an analog of a Control Law running on the Controller) to meet a certain Criterion (performance metric) expressed in terms of the distance from the desired state of the System.

In this view, we review a number of recent papers on assessment of resiliency, with two objectives in mind:

- First, to see whether or not the structural components of resiliency discussed in other papers fit into our model, which would serve the purpose of model validation.
- Second, to see whether or not a Performance Metric can be developed that would be useful in assisting in the assessment of resiliency.

The MITRE resilience review report [14] does not build any specific model of resilience, but introduces an interesting taxonomy composed of eight resilience categories. The categories differ regarding the ways how the resilience is implemented and their relation to the system components and events as presented in Figure 4. The taxonomy of resilience categories include: Adaptive Response, Deception, Detection/Monitoring, Dynamic Variations, Resilience Integrity, Isolation/Containment, Metrics/Assessment, and Cross-Area.

Rieger et al. [11] state that "resilience describes how systems operate at an acceptable level of normalcy despite disturbances or threats" and define explicitly a Resilient Control System as the one "that maintains state awareness and an accepted level of operational normalcy in response to disturbances, including threats of an unexpected and malicious nature." Thus, the definition is strictly consistent with the view presented earlier in this section. Among the specific issues to be considered when addressing the notion of resilience, the authors listed: latency, physical degradation, cyber security, and human performance.

Strigini [21] presents an interesting perspective on resilience, deriving the word from the Latin verb *resilire* (*re-salire*: to jump back), which literally means "the tendency or ability to spring back, and thus the ability of a body to recover its normal size and shape after being pushed or pulled out of shape, and therefore figuratively any ability to recover to normality after a disturbance." He confirms the technical meaning of the term referring "to materials recovering elastically after being compressed, and also in a variety of disciplines to designate properties related to being able to withstand shocks and deviations from the intended state and go back to a pre-existing, or a desirable or acceptable, state." The paper also confirms the approach presented here to building a model of resiliency, referring to feedback control and stability, as well as to fault tolerance and redundancy.

To summarize, the papers by Rieger et al. [11] and Strigini [21] address concepts directly compatible with those proposed when building the model of a resilient system in Figure 4. It contains a control system as an example of a cyberphysical system and includes all related components: threats that can exploit vulnerabilities in the controller, hazards/threats that may activate controller's faults, and a

hypothetical attack surface that consists of four interfaces through which the controller interacts with the world.

Thus, the model of a resilient system is nearly complete and can be viewed as validated. As indicated in the analysis of the MITRE report [14], an extension of the initial model is proposed, which includes an abstraction of an Attacker, capturing the essence of his actions, which is also illustrated in Figure 4.

V. Resiliency Metrics and Measures

A number of authors discuss various aspects of assessing resiliency, presenting numerous metrics and measures, using these terms interchangeably. For example, Strigini [21] discusses the entire array of measures related to quantitative reasoning about resilience (they should be in fact called metrics), including the following:

- measures of dependability in the presence of disturbances, which may be estimated empirically in operation or in a laboratory, or through probabilistic models (as functions of measures at component level)
- measures of the amount of disturbances that a system can tolerate, typically obtained from analyzing a system's design
- measures of probability of correct service given that a disturbance occurred ("coverage factors"), typically estimated empirically, often in a laboratory.

Additional measures for less technical categories of systems listed in [21] include:

- buffering capacity, which is essentially an "extent of tolerable disturbances";
- flexibility versus stiffness: the system's ability to restructure itself in response to external changes of pressure;
- margin: how closely or how precarious the system is operating relative to one or another kind of performance boundary;
- tolerance: how a system behaves near a boundary – whether the system gracefully degrades as stress/pressure increase or collapses quickly when pressure exceeds adaptive capacity;

In [17], the authors state that the "framework for resilience engineering is based on four key pillars: disruptions, system attributes, methods, and metrics," but do not create a more formal model of resilience beyond listing a number of components for each "pillar." The most interesting from our perspective are the Metrics, which include the following: time/cost to restore operation, time/cost to restore configuration (reconfigure), time/cost to restore functionality/performance, degree to which pre-disruption state is restored, potential disruption circumvented, and successful adaptations with time and cost constraints.

Almeida et al. [31] use a model similar to ours, but much less detailed, to reason about resilience of self-adaptive systems, calling the assessment process "benchmarking." They use several service related metrics, including:

- Performance: the number of operations the system is able to perform per unit time.
- Uptime: measure of the time the system is available during the benchmark procedure.
- Robustness: requires assessing the relative number of perturbations the system deals with gracefully, while maintaining system attributes values close to the desired specifications.

To better characterize self-adaptation capability, they consider other metrics that include:

- Time to react: the time elapsed from the exposure of the system to a perturbation until its recognition and decision to act upon.
- Time to adapt: the time necessary to execute the decided adaptation.
- Time to stabilize: the time the system takes to stabilize its operation.

Finally, stating that "as a system's ability to successfully adapt to perturbations depends on correctly deciding which perturbations to act upon, and doing it in a timely fashion," they include two additional metrics:

- Sensitivity: represents the ratio of adaptations performed to the number of perturbations submitted to the system.
- Degree of autonomy: portrays the system dependency on human operators.

On the other hand, Ramuhalli et al. [26] have a critical view of this approach to resilience metrics and state the following: "The bulk of these metrics are focused on system-level quantities (such as time to recover from an attack, percentage of available services, etc.). While these are important and help characterize the system performance, these are difficult to use for dynamic reconstitution, as computing such metrics in real-time (as the system is being reconstituted) from knowledge of only the configuration and/or connectivity is difficult." What they propose to use instead are indirect metrics and including graph metrics, "such as diameter, algebraic connectivity, average path length, clustering coefficient, although other graph statistics may be relevant and computable in real-time."

In an extensive report, Bodeau et al. [32] distinguish between two broad types of metrics relevant to cyber resiliency:

- Technical metrics, which evaluate the behavior of technologies and of technology dependent mission/business processes (particularly cyber defense processes);
- Organizational metrics, which evaluate organizational processes for resilience (in which cyber resiliency is – or should be – a consideration).

Both categories are, however, related to a much higher level of resiliency than that concerned in cyberphysical systems.

VI. Modeling Strategy

The essence of resilience is not to guarantee absolute protection against attacks or failures, but to provide successful recovery after disruptions. For example, Ramuhalli and his group at Pacific Northwest National Laboratory understand resilience as the degree of stability of the system at or near any operational state [33]. Similar views have been expressed by researchers at the Idaho National Laboratory [11] and others. Consequently, Vugrin et al., at SANDIA [34] state that "the cybersecurity community has voiced the opinion that cybersecurity strategies must expand beyond the protection-centric focus to incorporate cyber resilience principles." This has been advocated even earlier, by a national panel of researchers [35], stating that in case of a disruption such as an imminent security breach, what "cyberphysical systems require is either reconfiguration to reacquire the needed resources automatically or graceful degradation if they are not available."

In previous research, the authors have addressed this problem with respect to security [36]. An essential assumption in this approach was that a security breach may not necessarily cause complete system failure but just degradation of system services. The effects of a security breach were analyzed with respect to changes of system behavior in the following states: normal state, several degraded states (depending on the system or application), and failure state. The results led to a better understanding of consequences of such breaches and improvement of security policies.

The same strategy is applied in case of modeling resiliency. First, based on the model of resiliency developed in Section III, involving safety and security as essential resiliency components, an extended model is proposed involving an Attacker. Then, the Performance Metric can be used, which adequately reflects the distance between the Normal and Degraded states. Finally, a simulation tool is applied to evaluate resiliency under various scenarios.

The modeling process involves the Model-Based Environment, Möbius [37] which includes a number of modeling formalisms assisting in system performance and dependability modeling.

One of these formalisms involves Fault Trees that are widely used for modeling system safety property. An illustrative example of a car engine and wheels control, as a case of a cyberphysical system, is shown in Figure 5 [37], as an AND tree for potential engine failure, and can be enhanced by an OR tree for wheel failure. Running the simulator for a specific set of parameters constitutes an experiment, which results in calculating means and variances confirming specific hypotheses that can be related to safety evolving over time as a component of resiliency.

A newer modeling formalism, the Adversary VIew Security Evaluation (ADVISE) was developed recently to enhance Möbius and provide means for quantitative, state-based analysis of system security [38]. Building the ADVISE model relies on constructing an attack execution graph describing steps that an attacker might attempt to achieve specific goals. In addition, various attributes of the attacker are defined in his profile.

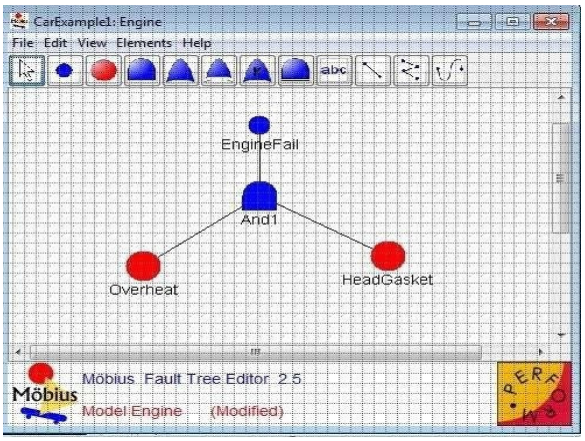


Fig. 5 Building Fault Trees in Möbius

Essential in the ADVISE model, the attack execution graphs (AEG) consist of attack step nodes, state variable nodes, and directed arcs between both types of nodes. State variable nodes store the state of a model during execution. During the run of an ADVISE model, the attacker (called an adversary) evaluates the state of the system, determines the most attractive attack step and attempts it. This decision process is repeated throughout the entire run of simulation.

ADVISE takes advantage of the Abstract Functional Interface (AFI) that facilitates the addition of new modeling formalism modules and new solver modules. Thanks to this feature, ADVISE models are designed to be composable with other Möbius models. It is anticipated that this capability can be used to combine security analysis in ADVISE with safety analysis using Möbius fault tree models for joint assessment of resiliency. Metrics for the assessment can be defined using the standard performance model available in Möbius, such as reward code expressions and impulse rewards. Specific metrics can be constructed to assess accomplishment of the goals by an attacker and risks associated with safety violations, to draw conclusions about resiliency levels.

VII. Conclusion

This paper addressed the assessment of resiliency of cyberphysical system in response to external disturbances, understood as hazards and threats causing safety and security violations, respectively, and related internal defects known as faults and vulnerabilities. A combined model for resiliency modeling and assessment was built, based on the view of feedback control theory enhanced with principles of

fault tolerance. This model was validated against the recent literature and enhanced with the view of potential attackers.

Resilience metrics were reviewed and analyzed by analogy with performance measures of the control system to assist in Design for Resilience. With the multitude of different approaches to resilience metrics and measures, it is suggested that those measures be selected, which best address the distance between the desired state of a system and the disrupted state level. The modeling strategy was proposed, based on using the Möbius modeling tool, which can address both security and safety issues as components of resiliency. Future work will involve combined simulations of fault-tree based (safety) and ADVISE models (security).

REFERENCES

[1] Zimmerman M.A., Resiliency Theory: A Strengths-Based Approach to Research and Practice for Adolescent Health, Health Education and Behavior, Vol. 40, No. 4, pp. 381– 383, August 2013.

[2] Holling C., Resilience and stability of ecological systems. Annual Review of Ecology and Systematics, Vol. 4, pp. 1-23, 1973.

[3] Christopher M., H. Peck, Building the Resilient Supply Chain. International Journal of Logistics Management, Vol. 15, No. 2, pp. 1-14, 2004.

[4] Adjetey-Bahun K. et al., A simulation-based approach to quantifying resilience indicators in a mass transportation system, Proc. ISCRAM2014, 11th Int'l Conference on Information Systems for Crisis Response and Management, University Park, Penn., May 18-21 2014.

[5] Goerger S.R., A.M. Madni, O.J. Eslinger, Engineered Resilient Systems: A DoD Perspective, Procedia Computer Science, Vol. 28, pp. 865-872, 2014.

[6] Castano V., I. Schagaev, Resilient Computer System Design, Springer-Verlag, Heidelberg, 2015.

[7] Suri N., G. Cabri (eds.), Adaptive, Dynamic, and Resilient Systems. CRC Press, Boca Raton, Fla., 2014.

[8] Hollnagel, E. Pariès, J. Woods, D. D. & Wreathall, J. (eds.). Resilience Engineering Perspectives. Vol. 3: Resilience Engineering in Practice. Ashgate, Farnham, UK, 2011.

[9] Ellison R. J. et al., Survivable network systems: An emerging discipline. Technical Report CMU/SEI-97-TR-013, Software Engineering Institute, Pittsburgh, Penn., 1997.

[10] Allen J., N. Davis, Measuring Operational Resilience Using the CERT® Resilience Management Model, Technical Note CMU/SEI-2010-TN-030. Software Engineering Institute, Pittsburgh, Penn., September 2010.

[11] Rieger C.G., K.L. Moore, T.L. Baldwin, Resilient Control Systems: A Multi-Agent Dynamic Systems Perspective. Proc. EIT 2013, IEEE International Conference on Electro/Information Technology, Rapid City, SD, May 9-11, 2013.

[12] Vugrin E.D., J. Turgeon, Advancing Cyber Resilience Analysis with Performance-based Metrics from Infrastructure Assessment. Int'l Journal of Secure Software Engineering, Vol. 4, No. 1, 2013.

[13] Alexander J.S., Achieving Mission Resilience for Space Systems. Spring 2012. URL: http://www.aerospace.org/2013/07/29/achieving-mission-resilience-for-space-systems/

[14] Pietravalle R., D. Lanz, Resiliency Research Snapshot. The MITRE Corporation, Bedford, Mass., June 2011.

[15] Bodeau D., R. Graubart, Cyber Resiliency Assessment: Enabling Architectural Improvement, Technical Report MTR120407, The MITRE Corporation. Bedford, Mass., May 2013.

[16] Caralli R.A. et al., CERT® Resilience Management Model, v1.0. Technical Report CMU/SEI-2010-TR-012. Software Engineering Institute, Pittsburgh, Penn., 2010.

[17] Madni A.M., S. Jackson, Towards a Conceptual Framework for Resilience Engineering, IEEE Systems Journal, Vol. 3, No. 2, pp. 181-191, June 2009.

[18] Teixeira A., Toward Cyber-Secure and Resilient Networked Control Systems. PhD Thesis, KTH Royal Institute of Technology, Stockholm, November 2014.

[19] Meyer, J. F. Defining and evaluating resilience: A performability perspective. Proc. PMCCC, Int'l Workshop on Performability Modeling of Computer and Communication Systems, Eger, Hungary, September 17-18, 2009.

[20] Bishop M. et al., Resilience Is More than Availability. Proc. NSPW'11, New Security Paradigms Workshop, Marin County, Calif., September 12-15, 2011, pp. 95–104.

[21] Strigini L., Fault Tolerance and Resilience: Meanings, Measures and Assessment, Resilience Assessment and Evaluation of Computing Systems, K. Wolter et al. (eds.), Springer-Verlag, Berlin, 2012.

[22] Athans M., P. Falb, Optimal Control. An Introduction to the Theory and Its Applications. McGraw-Hill, New York, 1966.

[23] Randell B. et al. (eds.), Predictably Dependable Computing Systems, Springer-Verlag, Berlin, 1995.

[24] Najjar W., J. Gaudiot, Network resilience: A measure of fault tolerance, IEEE Trans. Computers, Vol. 39, No. 2, pp. 174–181, February 1990.

[25] Axelrod W., Investing in Software Resiliency, CrossTalk: The Journal of Defense Software Engineering, Vol. 22, No. 6, pp. 20-25, September/October 2009.

[26] Ramuhalli P. et al., Towards a Theory of Autonomous Reconstitution of Compromised Cyber-Systems. Proc. HST2013, IEEE International Conference on Technologies for Homeland Security, Waltham, Mass. November 12-14, 2013.

[27] Ross R., J.C. Oren, M. McEvilley, Systems Security Engineering: An Integrated Approach to Building Trustworthy Resilient Systems. NIST Special Publication 800-160. National Institute of Standards and Technology, Gaithersburg, MD, May 2014.

[28] International Electrotechnical Vocabulary (IEV), International Electrotechnical Commission (IEC), Geneva, Switzerland. URL: http://www.electropedia.org/

[29] IEEE Software and Systems Engineering Vocabulary. IEEE Computer Society, Washington, DC, URL: http://computer.org/sevocab

[30] National Information Assurance (IA) Glossary. CNSS Instruction No. 4009. Committee on National Security Systems, 26 April 2010.

[31] Almeida R., H. Madeira, M. Vieira, Benchmarking the Resilience of Self-Adaptive Systems: A New Research Challenge. Proc. 29th IEEE Int'l Symposium on Reliable Distributed Systems, New Dehli, October 31 - November 3, 2010.

[32] Bodeau D., R. Graubart, L. LaPadula, P. Kertzner, A. Rosenthal, J. Brennan, Cyber Resiliency Metrics. Version 1.0, Rev. 1. Technical Report MP120053, The MITRE Corporation, Bedford, Mass. April 2012.

[33] Ramuhalli P., Theory of Resilience: A Framework for Resilient Design and Reconstitution of Cyber Systems, Project Flyer, Pacific Northwest National Laboratory, Richland, Wash., 2014. URL: http://cybersecurity.pnnl.gov/documents/projects/Theory_Flyer.pdf

[34] Vugrin E.D., R.C. Camphouse, Infrastructure resilience assessment through control design. International Journal of Critical Infrastructures, Vol. 7, No. 3, pp. 243-260, 2011.

[35] National Research Council, Committee for Advancing Software-Intensive Systems, Producibility Critical Code: Software Producibility for Defense, National Academies Press, Washington, DC, 2010.

[36] Kornecki A., J. Zalewski, W. Stevenson, Availability Assessment of Embedded Systems with Security Vulnerabilities, Proc. SEW-2011, 34th IEEE Software Engineering Workshop, Limerick, Ireland, June 20-21, 2011, pp. 42-47.

[37] Möbius: Model-Based Environment for Validation of System Reliability, Availability, Security and Performance. Performability Engineering Research Group, University of Illinois, Urbana-Champaign, Ill., 2014. URL: https://www.mobius.illinois.edu/

[38] Ford M.D. et al., Implementing the ADVISE Security Modeling Formalism in Möbius. Proc. DSN '13, 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Budapest, Hungary, June 24-27, 2013.G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

5th Workshop on Advances in Programming Languages

PROGRAMMING languages are programmers' most basic tools. With appropriate programming languages one can drastically reduce the cost of building new applications as well as maintaining existing ones. In the last decades there have been many advances in programming languages technology in traditional programming paradigms such as functional, logic, and object-oriented programming, as well as the development of new paradigms such as aspect-oriented programming. The main driving force was and will be to better express programmers' ideas. Therefore, research in programming languages is an endless activity and the core of computer science. New language features, new programming paradigms, and better compile-time and run-time mechanisms can be foreseen in the future.

The aims of this event is to provide a forum for exchange of ideas and experience in topics concerned with programming languages and systems. Original papers and implementation reports are invited in all areas of programming languages.

Topics

Major topics of interest include but are not limited to the following:

-
- Automata theory and applications
- Compiling techniques
- Domain-specific languages
- Formal semantics and syntax
- Generative and generic programming
- Grammarware and grammar based systems
- Knowledge engineering languages, integration of knowledge engineering and software engineering
- Languages and tools for trustworthy computing
- Language theory and applications
- Language concepts, design and implementation
- Markup languages (XML)
- Metamodeling and modeling languages
- Model-driven engineering languages and systems
- Practical experiences with programming languages
- Program analysis, optimization and verification
- Program generation and transformation
- Programming paradigms (aspect-oriented, functional, logic, object-oriented, etc.)
- Programming tools and environments
- Proof theory for programs
- Specification languages
- Type systems
- Virtual machines and just-in-time compilation
- Visual programming languages

Steering Committee

Janousek, Jan, Czech Technical University, Czech Republic

Luković, Ivan, University of Novi Sad, Serbia

Mernik, Marjan, University of Maribor, Slovenia

Slivnik, Boštjan, University of Ljubljana, Slovenia

Event Chair

Porubän, Jaroslav, Technical University of Kosice, Slovakia

Program Committee

Barisic, Ankica, Universidade Nova de Lisboa, Portugal

Horvath, Zoltan, Eotvos Lorand University, Hungary

Janousek, Jan, Czech Technical University, Czech Republic

João Varanda Pereira, Maria, Instituto Politecnico de Braganca, Portugal

Kardaş, Geylani, Ege University International Computer Institute, Turkey

Kollár, Ján, Technical University of Kosice, Slovakia

Kosar, Tomaž, University of Maribor, Slovenia

Liu, Shih-Hsi Alex, California State University, United States

Luković, Ivan, University of Novi Sad, Serbia

Mandreoli, Federica, University of Modena, Italy

Martínez López, Pablo E. "Fidel", Universidad Nacional de Quilmes, Argentina

Mernik, Marjan, University of Maribor, Slovenia

Milasinovic, Boris, University of Zagreb Faculty of Electrical Engineering and Computing, Croatia

Moessenboeck, Hanspeter, Johannes Kepler Universitat Linz, Austria

Papaspyrou, Nikolaos, National Technical University of Athens, Greece

Rangel Henriques, Pedro, Universidade do Minho, Portugal

Sierra Rodríguez, José Luis, Universidad Complutense de Madrid, Spain

Slivnik, Boštjan, University of Ljubljana, Slovenia

Splawski, Zdzislaw, Wroclaw University of Technology, Poland

van der Meer, Arjan, Eindhoven University of Technology

Watson, Bruce, Stellenbosch University, South Africa

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 117–122

DOI: 10.15439/2015F392
ACSIS, Vol. 6

Using the Interaction Flow Modelling Language for Generation of Automated Front–End Tests

Karel Frajták, Miroslav Bureš, Ivan Jelínek
Department of Computer Science
Faculty of Electrical Engineering
Czech Technical University
Karlovo nám. 13, 121 35 Praha 2, Czech Republic
Email: {frajtak, buresm3, jelinek}@fel.cvut.cz

***Abstract*—In the paper we explore the possibilities of automated test-case generation from the IFML model of application front–end. As opposed to the previous core UML standard, IFML captures the structure and properties of the application user interface, which gives us new possibilities in model–based test case generation: produced test cases have a higher probability of being consistent and of respecting the real feasibility of the tests in the tested application. In the presented solution we leverage the capabilities of an IFML model to capture details of front–end components to generate front–end automated tests, exercising particular actions in the tested application front–end to verify its expected behaviour according to an IFML model. The approach is based on the transformation of an IFML model to an application front–end test model — a more straightforward structure for the automated generation of test cases. Then, based on the defined rules, the abstract test cases are created from the model. The abstract test cases are then transformed using a template engine, to particular physical automated test cases which can be run to test the application.**

I. Introduction

TODAY'S efficiency and short time–to–market is the key factor in software development that creates pressure on software development teams and a demand for more efficient methods of software development and testing.

The mission of the development is to deliver the system (or a modification of the system) in an agreed time. In the iterative development it means to fix the issues from the previous cycle, to add new features and to test the system and to verify that every feature of the software works according to the specification. What if a customer suddenly wants a new feature that the developers will have a hard time implementing? Or what if this new functionality affects the entire application, so regression effect rates of the code changes are high and the reliability of previously stable parts of the application is challenged?

Testing in such a cycle is often challenging. Preparation and execution of the tests, if performed manually, requires time which is often not available. This can also affect the accuracy of prepared test cases.

Automation of the test cases is one of the possible ways of how to make the process more efficient.

In this paper we are proposing a model–driven approach to front–end web application testing based on the Interaction Flow Modelling Language (IFML, [3]). We are going to describe the process of transforming the IFML model of the tested application to a set of front–end test cases. The automatic generation without of these tests from the IFML model guarantees their consistency.

In this field, UML [11] is a widely adopted modelling language made to visualize the design of the system. Nevertheless, UML does not capture all aspects of the application. One area where UML is lacking vocabulary and tools is in the modelling of the user interface and interaction. To overcome this gap, a Web Modelling Language (WebML [15]) was created introducing visual notations and a methodology for designing complex data–intensive Web applications. This language later evolved into IFML to cover a wider spectrum of front–end interfaces and the data flows between the application front–end components. IFML was later adopted by the Object Management Group (OMG, [16]) as an industrial standard.

II. Problem description

During the development stage, changes are frequently made to the web application code base. On the front–end the page layout can change, input elements are added or removed, data–flow of the pages is modified. All of these changes must be tested in order to prove that no error was introduced and that everything works as expected. Without the model-driven approach, both the code of the application and the tests are created manually. Every change made to the code must be synchronized with the tests, so that the tests are testing new functionality with new input elements and new corner case input values. When an element is added to a form, all functional tests associated with that form must be modified accordingly. This maintenance of the automated test scripts causes significant overhead in the development of the software project.

III. Potential of IFML for test case generation

Data driven application front–end is usually built using reusable components (forms, list views, detail views, etc.). These components have expected behaviour. For example, forms are placed on the page to be filled in with data and sent to the server, lists show record details for the user to view or allow him or her to select one or more records and perform

an action on these. All of these operations can be modelled using the IFML notation (see an example in Figure 1).

With precise models and proper tooling (for instance IBM Rational Software Architect, Enterprise Architect, AndroMDA) these models can be transformed to code, different models or just to generate the system scaffolding. Developing an application with tens of screens with various components (forms, list views, etc.) can be a lengthy and repetitive task — every form and every list view must be manually created. This process leads to copy paste style of programming and any possible defects can be easily cloned and introduced many times in the application. Even with the use of a user interface component framework this can be a problem. The efficiency of the process can be increased by the generation of the user interface (UI) from the model [4, 5]. While this is efficient — it is certainly easier to create a model of a number of components and their interaction than to implement them physically — it still does not prove that the application is error free and can be delivered to the users. And in most of the cases, we don't have resources to test manually every screen in the application whenever there's a code change.

A similar approach can be also used to generate test case scenarios — we know how to test the basic functionality of a component and what the code for such a test should look like. For this reason it is highly recommended to use the IFML model originally used for the front–end code generation to generate test case scenario code just by using a different template.

IV. Related work

Although WebML has been used for more than ten years, IFML is relatively new and was recently standardized. The first applications of the standard are emerging, for instance, a systematic model–driven reverse engineering process to generate an IFML representation from such applications is presented in [17].

IFML notation can be easily extended by adding new containers, components, events or by applying custom UML stereotypes to them as described in [4]. The authors added new components and events (swipe, camera event, location sensor event) to be able to describe mobile specific interfaces and interaction.

IFML represents a prospective modelling tool to describe application front–end and a flexible and easily extensible notation. Hence, we decided to use its capabilities to generate front–end test case scenarios.

In the previous approaches, a general purpose modelling languages, such as UML, that described the system high–level model were often used for system code generation. UML models have normally been used for the process of automated code generation from the model, for example [12, 1, 13]. The same applies for the generation of test cases from the tested application model.

From the previous approaches, sequence diagrams [2, 19], state chart [10] or activity diagrams [9, 14] are used, but these diagrams and the models they represent are more focused on describing the application structure or data flow not the user front–end interaction. In [19] sequence diagrams as the most suitable for precise and detailed description of a system's actions and behaviour. UML notation was also used to generate the user interface. In [5] a new diagram called user interface diagram was introduced with a user interface specialization.

Our goal is to use the approach of generating the test cases from the application model, but instead of UML, which is already covered in the previous work, we are going to generate automated test cases from the IFML model. This area is currently lacking sufficient theoretical support since IFML is a quite new modelling language. In [4] the proposals were verified using manual testing by testers.

V. Proposed solution

Our proposal of generating automated end–to–end test case scenarios from IFML model is based on the set of transformations, which are outlined in Figure 2.

In the proposed solution, the XML representation of an IFML model is converted into a front–end model using a predefined set of rules. The model is then used to generate abstract test case scenarios. Physical details are added to these scenarios by templates and a set of executable test case scenarios are created. Details of this process are following further on.

VI. Front–end model

In this section we present details of the front–end model, used for the transformation process introduced above. The aim of this model is to formalize the user interaction with the application.

We consider this model more suitable for generation of the test cases, as the IFML notation is too rich and descriptive for our use. In this proposal, we have used an already defined and published formal model, verified in our previous work [7, 6].

We define a view window W as a set of view containers and a view container K as a tuple $\langle C, N, A, E, D, M\rangle$, where

- C is a hierarchical set of view components placed into this view container (components can be nested)
- N is a set of navigation flows defined as $N : C \cup E \cup D \cup B \rightarrow C \cup A$. The user triggers an event E on view component C with data–bound by D resulting in displaying another view component (or the same one) or triggering an action A
- A is a set of actions executed prior to updating the state of the user interface
- E is a set of events a view container and a view component is associated with, the effect of event is the interaction flow
- B is a set of data bound variables whose values will be used in navigation flow
- D is set of data binding expressions d defined as $d : C \rightarrow B$, these expressions extract a value from a view component, for example it describes how to get the numeric value from a text input field

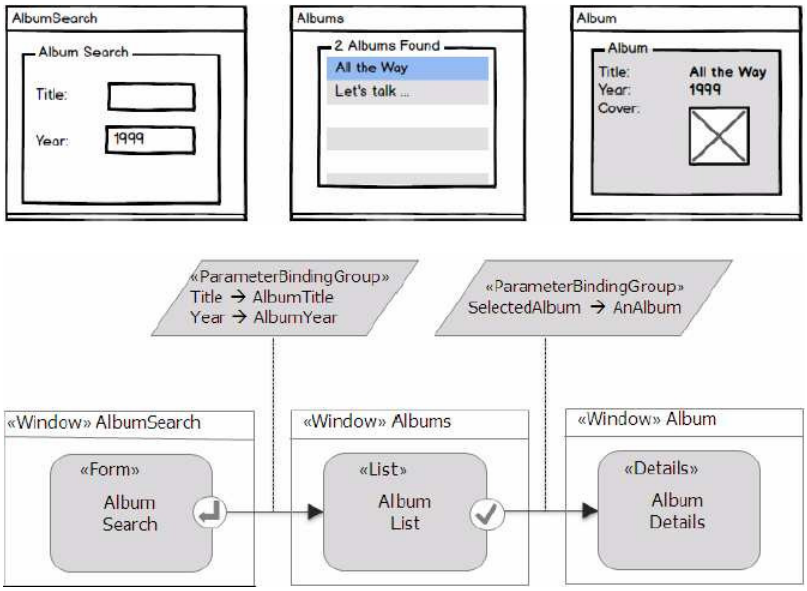


Fig. 1. Example of an IFML model of a "Search for an album" use case

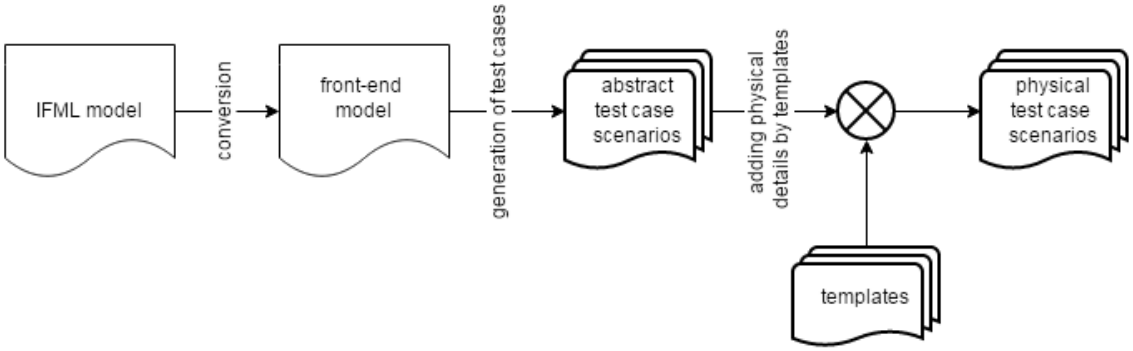


Fig. 2. Transformation of an IFML model to physical automated test case scenarios

- M is a set of custom metadata m to describe the meta–properties of the view components defined as $m : C \rightarrow GM$, where GM is an extensible global set of model meta–properties holding IFML model properties, UML stereotypes and other custom properties

In our example introduced above, the front–end model for the view window from our album search example can be described as the following (we have omitted the definition of K_{Albums} for brevity): $W = \{K_{AlbumsSearch}, K_{Albums}\}$

$K_{AlbumsSearch} = \{$
$C = \{AlbumSearchForm = \{AlbumTitle, AlbumYear\}\}$
$N = \{AlbumSearchForm \cup Submit \cup \{Title, Year\} \rightarrow Albums\}$
$A = \emptyset$
$E = \{Submit\}$
$D = \{AlbumTitle \rightarrow Title, AlbumYear \rightarrow Year\},$
$B = \{Title, Year\},$
$M = \{AlbumSearchForm \rightarrow \{Form\}, AlbumTitle \rightarrow \{SimpleField, String\},$
$AlbumYear \rightarrow \{SimpleField, Year\}\}$
$\}$

The abstract test case scenario T is defined as $T : N \cup C \cup V \rightarrow 0, 1$, N and C are defined as before and V is a set of rules that all has to be matched for the test not to be marked as failed.

For simplicity just assume that when the search operation has finished, the albums view container is displayed (no matter how many results it will show). In that case the result of T is success when C is $AlbumList$:

$$T = 1 \Leftrightarrow V = \{Equals(C, AlbumList)\}$$

For our purposes the front–end model is serialized into JSON format, which is easy to read and can be injected directly into the template engine.

VII. The transformation process

As we have already introduced, the first step of the process is the transformation of an IFML model into a front–end model. Modelling tools use XML format to persist the models

(see Listing 1), however this format is quite heavy for further processing. We have decided to transform the XML format into more readable and lightweight JSON form. For this conversion process, we have defined a set of rules.

```
<interactionFlowModelElements name="AlbumList" xsi:
    type="ext:IFMLWindow">
 <viewElements xsi:type="ext:Form"
   name="AlbumSearchForm">
  <viewElementEvents xsi:type="ext:OnSubmitEvent"
      name="Search">
   <outInteractionFlows>
    <parameterBindingGroup .../>
   </outInteractionFlows>
  </viewElementEvents>
  <viewComponentParts xsi:type="core:DataBinding"
      name="Album"/>
  <viewComponentParts xsi:type="ext:SimpleField"
      name="Title"/>
  <viewComponentParts xsi:type="ext:SimpleField"
      name="Year"/>
 </viewElements>
</interactionFlowModelElements>
```

Listing 1. XML representation of an IFML model

For each IFML element in order to be used in our model a rule is defined. For example for an IFML element with the name “viewElements” or “viewComponentParts” an entry is added to the “components” array, for every “xsi:type” attribute with the value ”ext:Form” a “type: form” entry is added to the “metadata”. In order to track the original elements we also copy the element id that uniquely identifies it. The rule is a simple JavaScript method matching properties of an XML element. For our example presented in Figure 1, the respective XML notation of the IFML model is given in Listing 1 and the JSON description of the front–end model in Listing 2.

```
{
  "name": "AlbumList",
  "metadata": [{ "type": "window" }],
  "components": [{
    "name": "AlbumSearchForm",
    "metadata": [{ "type": "form" }],
    "variables": [{
      "name": "Year",
      "metadata": [{
        "dataType": "int",
        "constraint": "not-empty"
      }]
    },{
      "name": "Title",
      "metadata": [{
        "dataType": "int",
        "constraint": "year"
      }]
    }],
    "binding": [{
      "name": "Title",
      "from": "AlbumTitle"
    },{
      "name": "Year",
      "from": "AlbumYear"
    }],
    "events": [{
      "type": "submit",
      "target": "Albums"
    }],
    "components": [{
      "name": "AlbumTitle",
      "metadata": [{ "type": "field" }]
    },{
      "name": "AlbumYear",
      "metadata": [{ "type": "field" }]
    }]
  }]
}
```

Listing 2. XML representation of an IFML model

In the next step the front–end model is transformed into a set of abstract test case scenarios. These scenarios are independent of the specific technological platform or programming language used to implement it — using the abstracted test cases gives us flexibility to generate the test case in different scripting languages and test automation APIs. From a technical perspective the JSON representation of the front–end model is transformed with the use of predefined transformation rules to JSON representation of the abstract test case scenario (see Listing 3).

```
{
  "name": "AlbumSearchForm",
  "id": "65de40fd-8283",
  "specs": [
    {
      "name": "Search",
      "type": "search",
      "steps": [
        "fill": {
          "locator": "Year",
          "type": "year"
        },
        "submit": { "locator": "AlbumSearchForm" },
        "waitFor": { "locator": "AlbumList" }
      ]
    },
    {"name":"Reset" ...}
  ]
}
```

Listing 3. Abstract test case scenario

In the last step the abstract test case scenarios are transformed into executable (platform specific or programming language specific) test case scenarios. In this process a template system is used to generate the test cases. The template system chooses the desired template from the templates library (WebdriverIO [21] template was used to generate code for our example). The template system is based on the JavaScript Underscore template [18] capability. Abstract test case scenario serialized into JSON format is supplied as a parameter and then processed by the template engine (see Listing 4).

```
var webdriverio = require('../index');
var templates = require('/templates');

describe('<%=scenario.name%>', function() {
  var client = {};
  jasmine.DEFAULT_TIMEOUT_INTERVAL = 9999999;

  beforeEach(function() {
    client = webdriverio.remote({
      desiredCapabilities:{
        browserName: 'phantomjs'}
    });
    client.init();
  });

<% _.each(scenario.specs, function(spec) { %>
  <% var specTemplate = specTemplates.getTemplate(
```

```
      spec.type); %>
  it('<%=toSpecName(spec.name)%>', function(done) {
    <%=specTemplate.renderTemplate(spec); %>
  });
<%});%>

  afterEach(function(done) { client.end(done); });
});
```

Listing 4. Underscore template

In Listing 5 we present the result of the transformation for example from Figure 1. The executable JavaScript is code created to be executed by Jasmine [8] test runner. In the example, we test the basic functionality of a search form. The test describes 'AlbumSearchForm' test suite with a spec 'should search' (spec is named set of expectations to be met). The spec function makes call to a browser automation tool Selenium WebDriver [18] via binding library WebdriverIO. If we want to use different programming language to implement the test cases, a different template can be used to generate physical test case scenarios from abstract test case scenarios.

```
var webdriverio = require('../index');
var templates = require('/templates');

describe('AlbumSearchForm', function() {
  var client = {};
  jasmine.DEFAULT_TIMEOUT_INTERVAL = 9999999;
  beforeEach(function() {
    client = webdriverio.remote({
      desiredCapabilities:{
        browserName: 'phantomjs'}
    });
    client.init();
  });

  it('should search', function(done) {
    client.url('...')
      .setValue('#Year', '2015')
      .submitForm('#AlbumSearchForm')
      .waitForExist('#AlbumList', 2000);
  });

  afterEach(function(done) { client.end(done); });
});
```

Listing 5. Code generated for AlbumSearchForm form from the IFML model (some code left for brevity)

The code first initializes the Selenium web driver to be used with windowless browser called PhantomJS, then navigates to our application page, fills value 2015 into the Year input field, submits the form and waits for AlbumList element to appear. If the element is not displayed within 2 seconds, the test fails.

VIII. Verification

The proposed solution is currently in the implementation stage with the first results arising from experiments. We have collected the initial feedback from the users of our prototype and adjusted the model accordingly. We have been experimenting with a set of IFML models created for 3 applications we created for our needs and the results are promising. The test case scenarios were correctly generated, but in some specific cases, the generation process should be improved upon further, which nevertheless represents an implementation task. As mentioned before the added value is the automatic generation of test cases when any change is made to the IFML model.

Our solution quickly discovered problems in tested applications in the cases when

- changes were made to existing server–side code introducing an error in processing the client data,
- changes were made to existing client side code introducing an error in JavaScript components,
- new elements were added with faulty behaviour or
- new elements were added, but the values of these components were not handled properly when sent to server.

On several occasions, we ended up in a situation where some tests were generated syntactically correctly but the semantics of expected result assertions was not corresponding to the state of the tested application. This would often lead to the IFML model, a conversion rule or a used template being updated. It was a means of feedback and an indication of how to evolve the solution to be fully functional. Only occasionally we have had to remove an IFML feature so that the test suite could be generated.

The biggest advantage of the use of an IFML model instead of an UML model as a base model for test case generation is the power of IFML to describe the front–end views, components and interaction between them. The test cases are therefore potentially more consistent.

IX. Conclusions and future work

IFML is a relatively new notation recently standardized by OMG. So far it has not been widely adopted; currently there are only 2 tools on the market (commercial WebRatio [22] and open source Eclipse plugin [20]). The primary target of IFML is to express the content, user interaction and control behaviour of the front–end of software applications, which are some of the key aspects of the application.

The first results of experiments with our proposed solution show that using an IFML model for test case generation is viable and promising. Initial feedback from experiments with our solution confirms an advantage of IFML — it does describe the front–end directly which gives relative high level of assurance about the precision of the generated test case scenarios and their ability to be executed.

In the future work, we are going to improve and extend the proposed solution to support more IFML constructs and be able to help in large–scale development projects. Our solution does not currently support any components other than forms and list views, but new UML stereotype can be added, or the IFML notation can be extended by the addition of a new component to change the transformation and generate new types of test case scenario.

Acknowledgment

This research has been supported by MŠMT under research program No. 6840770014 and by Grant Agency of the CTU in Prague under grant SGS14/076/OHK3/1T/13.

REFERENCES

[1] Manoli Albert et al. "Automatic generation of basic behavior schemas from UML class diagrams". English. In: *Software & Systems Modeling* 9.1 (2010), pp. 47–67. ISSN: 1619-1366. DOI: 10.1007/s10270-008-0108-x. URL: http://dx.doi.org/10.1007/s10270-008-0108-x.

[2] A. Bandyopadhyay and S. Ghosh. "Test Input Generation Using UML Sequence and State Machines Models". In: *Software Testing Verification and Validation, 2009. ICST '09. International Conference on*. Apr. 2009, pp. 121–130. DOI: 10.1109/ICST.2009.23.

[3] Marco Brambilla and Piero Fraternali. *Interaction Flow Modeling Language: Model-Driven UI Engineering of Web and Mobile Apps with IFML*. Morgan Kaufmann, 2014.

[4] Marco Brambilla, Andrea Mauri, and Eric Umuhoza. "Extending the Interaction Flow Modeling Language (IFML) for Model Driven Development of Mobile Applications Front End". English. In: *Mobile Web Information Systems*. Ed. by Irfan Awan et al. Vol. 8640. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 176–191. ISBN: 978-3-319-10358-7. DOI: 10.1007/978-3-319-10359-4_15. URL: http://dx.doi.org/10.1007/978-3-319-10359-4_15.

[5] F. Ferri. *Visual Languages for Interactive Computing: Definitions and Formalizations*. Premier reference source. Information Science Reference, 2008. ISBN: 9781599045368. URL: https://books.google.co.uk/books?id=LNOSq-q7wfoC.

[6] Karel Frajták, Miroslav Bureš, and Ivan Jelínek. "Formal specification to support advanced model based testing". In: *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*. IEEE. 2012, pp. 1311–1314.

[7] Karel Frajták, Miroslav Bureš, and Ivan Jelínek. "Manual testing of web software systems supported by direct guidance of the tester based on design model". In: *World Academy of Science, Engineering and Technology* (2011), pp. 542–545.

[8] *Jasmine, behavior-driven development framework for testing JavaScript code @ONLINE*. http://jasmine.github.io.

[9] S. Kansomkeat, P. Thiket, and J. Offutt. "Generating test cases from UML activity diagrams using the Condition-Classification Tree Method". In: *Software Technology and Engineering (ICSTE), 2010 2nd International Conference on*. Vol. 1. Oct. 2010, DOI: 10.1109/ICSTE.2010.5608913.

[10] Supaporn Kansomkeat and Wanchai Rivepiboon. "Automated-generating Test Case Using UML Statechart Diagrams". In: *Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*. SAICSIT '03. Johannesburg, South Africa: South African Institute for Computer Scientists and Information Technologists, 2003, pp. 296–300. ISBN: 1-58113-774-5. URL: http://dl.acm.org/citation.cfm?id=954014.954046.

[11] Andrey Karpov. *Myths about static analysis. The third myth - dynamic analysis is better than static analysis @ONLINE*. http://www.viva64.com/en/b/0117/. Accessed: 2013-09-04. Nov. 2011.

[12] D. Kundu, D. Samanta, and R. Mall. "Automatic code generation from unified modelling language sequence diagrams". In: *Software, IET* 7.1 (Feb. 2013), pp. 12–28. ISSN: 1751-8806. DOI: 10.1049/iet-sen.2011.0080.

[13] Abid Mehmood and Dayang N.A. Jawawi. "Aspect-oriented model-driven code generation: A systematic mapping study". In: *Information and Software Technology* 55.2 (2013). Special Section: Component-Based Software Engineering (CBSE), 2011, pp. 395–411. ISSN: 0950-5849. DOI: http://dx.doi.org/10.1016/j.infsof.2012.09.003. URL: http://www.sciencedirect.com/science/article/pii/S0950584912001863.

[14] Chen Mingsong, Qiu Xiaokang, and Li Xuandong. "Automatic Test Case Generation for UML Activity Diagrams". In: *Proceedings of the 2006 International Workshop on Automation of Software Test*. AST '06. Shanghai, China: ACM, 2006, pp. 2–8. ISBN: 1-59593-408-1. DOI: 10.1145/1138929.1138931. URL: http://doi.acm.org/10.1145/1138929.1138931.

[15] N. Moreno, P. Fraternali, and Antonio Vallecillo. "WebML modelling in UML". In: *Software, IET* 1.3 (June 2007), pp. 67–80. ISSN: 1751-8806.

[16] *Object Management Group @ONLINE*. http:///www.omg.org.

[17] Roberto Rodriguez-Echeverria et al. "IFML-based Model-Driven Front-End Modernization". In: (2014).

[18] *Selenium, web browser automation @ONLINE*. http://www.seleniumhq.org.

[19] Aristos Stavrou and GeorgeA. Papadopoulos. "Automatic Generation of Executable Code from Software Architecture Models". English. In: *Information Systems Development*. Ed. by Chris Barry et al. Springer US, 2009, pp. 1047–1058. ISBN: 978-0-387-78577-6. DOI: 10.1007/978-0-387-78578-3_36. URL: http://dx.doi.org/10.1007/978-0-387-78578-3_36.

[20] *The open source IFML editor - Based on Sirius @ONLINE*. https://github.com/ifml/ifml-editor.

[21] *WebdriverIO, Selenium 2.0 bindings for NodeJS @ONLINE*. http://webdriver.io.

[22] *WebRatio @ONLINE*. http://www.webratio.io.

2nd Workshop on Emerging Aspects in Information Security

ADMITTEDLY,information security works as a backbone for protecting both user data and electronic transactions. Protecting the communication and data infrastructure of an increasingly inter-connected world has become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The Emerging Aspects in Information Security (EAIS'15) workshop focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The workshop is an umbrella for all information security technical aspects. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. EAIS'15 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the EAIS'15 workshop can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing to elaborate integrated security solutions, e.g. integrate different risk-based management systems
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas
- To present latest security challenges, especially with respect to EC Horizon 2020

Topics

Topics of interest include but are not limited to:

-
- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis
- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Individual identification and privacy protection
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management in different application domains
- Tools supporting security management and development
- Emerging technologies and applications
- Digital forensics and crime science
- Misuse and intrusion detection
- Security knowledge management
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

Event Chairs

Awad, Ali Ismail, Luleå University of Technology, Sweden

Bialas, Andrzej, Institute of Innovative Technologies EMAG, Poland

Program Committee

AbdAllah, Mohamed Mostafa, Yanbu Industrial College, Saudi Arabia

Bun, Rostyslav, Lviv Polytechnic National University

Clarke, Nathan, Plymouth University, United Kingdom

Cyra, Łukasz, European Commission - Joint Research Centre Institute for the Protection & Security of the Citizen

Dworzecki, Jacek, Police Academy in Szczytno

Fernandez, Eduardo B., Florida Atlantic University, United States

Furnell, Steven, Plymouth University, United Kingdom

Furtak, Janusz, Military University of Technology, Poland

Geiger, Gebhard, Technical University of Munich, Faculty of Economics

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Hämmerli, Bernhard M., Hochschule für Technik+Architektur (HTA), Switzerland

Hasssaballah, M., South Valley University, Egypt

Kalbarczyk, Zbigniew, University of Illinois at Urbana-Champaign

Kapczynski, Adrian, Silesian University of Technology, Poland

Klamka, Jerzy, Polish Academy of Sciences

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 125–131

DOI: 10.15439/2015F140
ACSIS, Vol. 6

Fully Homomorphic Encryption for Secure Computations in Protected Database

Darya Chechulina, Kirill Shatilov, Sergey Krendelev
Department of Information Technology, Novosibirsk State University,
Novosibirsk, Russia
Email: chechulina, shatilov, krendelev@ccfit.nsu.ru

***Abstract*—Outsourced computations and, more particularly, cloud computations, are widespread nowadays. That is why the problem of keeping the data security arises. Multiple fully homomorphic cryptosystems were proposed in order to perform secret computations in untrusted environments. But most of the existent solutions are practically inapplicable as they require huge computation resources and produce big (∼1Gb) keys and ciphertexts. Therefore, we propose the undemanding fully homomorphic scheme with practically acceptable (∼few Kb) keys and output data. Our solution uses modular arithmetic in order to avoid the increase in data size. We have validated our approach through the implementation of the proposed cryptosystem. The details of used algorithms and the results of security evaluation are covered in this paper.**

I. Introduction

NOWADAYS Information Technologies and, particularly, computations over various data are the important part of our living and business processes. Modern trend to outsource computations to third-parties has aroused a problem of keeping the security of one's data. Cloud computing and other cases of giving the access to the personal data are affected by threat of exposing vulnerable data to unauthorized parties. Using a fully homomorphic encryption (FHE) scheme in secure computations helps to avoid the data leakage.

Originally a conception of FHE was introduced by Rivest, Adleman and Dertouzous in their paper [5]. Since people wanted to be able to perform the computations over the encrypted data, the problem of privacy homomorphism became very actual one in cryptography at whole. The first attempt of proposing FHE scheme belongs to Gentry [1]. After publication of the scheme's idea he introduced the implementation of his algorithm in conjunction with Halevi [2]. Then a lot of improvements of Gentry's work were proposed. But all of them were criticized as they required significant computing resources due to the usage of complex mathematical tools and produced big sizes of keys and output data [6].

Most of the proposed encryptions suffer from inefficiency to the practical use; therefore the problem of computations security is still actual [4]. That is why our ultimate goal for FHE developing is researching for the previously not used but efficient mathematical technics to make practical implementation. As a result, we introduce a new fully homomorphic scheme that doesn't require massive computation resources and provide acceptable sizes of encryption keys and output values.

This research was performed in Novosibirsk State University under support of the Ministry of education and science of Russia (contract no. 02.G25.310054)

The next section of this paper features the mathematical bases of our approach as it gives some fundamental definitions. Section 3 describes the properties of computations over encrypted data. After it, in Section 4, we show the core components of the proposed fully homomorphic encryption cryptosystem. Section 5 covers the evaluation of our scheme's security. Then, Section 6 discusses possible applications of the developed homomorphic encryption including the implemented one and summarizes our achievements.

II. Mathematical Foundation

This section gives essential mathematical bases. Let us discuss what a fully homomorphic cryptosystem means. Formally, such a scheme allows performing computation over the encrypted data without their decryption. In other words, an encryption algorithm E and a decryption D should satisfy the following conditions:

$$c_1 = E(a_1), c_2 = E(a_2)$$

$$D(f(c_1, c_2)) = f(a_1, a_2)$$

where c_1, c_2 are the ciphertexts and f is an arbitrary, efficiently computed function. In order to avoid the increase in data size we use modular arithmetic in our approach. Thus, the main idea of the proposed solution is as follows: we have a set of relatively prime numbers $(m_1, m_2, \ldots, m_k)$. The plaintext P we associate with the set $P = (P_1, P_2, \ldots, P_k)$ where $P_i = P \ mod \ m_i, i = 1, \ldots, k$. This set is encrypted with proposed algorithm that will be described in details in the following section. The approach has the only constraint: the result of all the mathematical operations can't exceed the number $M = m_1 \cdot \ \ldots \ \cdot m_k$.

Firstly we consider the simplified algorithm for the ring Z_m and the modulus m only. To encrypt P we select a secret vector $x = (x_1, \ldots, x_n), x_i \in Z_m$. Then we construct a vector $a = (a_1, \ldots, a_n), a_i \in Z_m$ as follows:

$$(a, x) = P \ mod \ m$$

It is worth noting that in general every number m can be represented as $m = p_1^{\alpha_1} \cdot \ \ldots \ \cdot p_s^{\alpha_s}$, where $p_1, \ldots, p_s$ are prime numbers. Thus, it is enough to make all the necessary steps

of the algorithm for the power of prime number only, or, in the simplest case, only for the prime number.

So, let m be a prime number. Now we will describe some mathematical details of the proposed approach.

The scalar product can be considered as a linear function:

$$h(x_1, \ldots, x_n) = (u, x), u, x \in Z_m^n$$

$$x = (x_1, \ldots, x_n)$$

$$u = (u_1, \ldots, u_n)$$

Thus, a linear function is completely determined by the vector u.

The secret point is defined as a vector $x = (x_1, \ldots, x_n)$. Thereby, to represent the number P, we must construct a vector $v \in Z_m^n$ as follows:

$$(v, x) = P \ mod \ m$$

This task belongs to the standard linear algebra and can be easily solved. Thus, v is called a ciphertext for P.

III. Computations over Encrypted Data

As it was previously mentioned, the proposed encryption allows performing the computations over ciphertexts.

A. Addition

Addition of vectors is equivalent to addition of their components with given modulus m according to the properties of the scalar product and the modular arithmetic. So, if we have representations for two numbers P_1 and P_2:

$$(v, x) = P_1 \ mod \ m$$

$$(u, x) = P_2 \ mod \ m$$

and in general the sum of the simple linear functions is defined as:

$$(v, x) + (u, x) = (v + u, x)$$

thus:

$$(v + u, x) \ mod \ m = [(v, x) + (u, x)] \ mod \ m =$$

$$= P_1 \ mod \ m + P_2 \ mod \ m = (P_1 + P_2) \ mod \ m$$

Let us note that addition keeps the size of vectors. That means the resulting vector has the same length as the initial ciphers.

B. Multiplication

Multiplication of vectors v and u in common way leads to the increase in the result's length almost n times:

$$w = v \cdot u = (v_1 u_1, v_1 u_2, \ldots, v_n u_n)$$

In order to prevent vectors' length growth, we define a specific kind of multiplication. Let the secret vector satisfy the following condition:

$$x_i x_j = \sum_{k=1}^{n} \gamma_{ijk} x_k \ mod \ m \tag{1}$$

Also generally, the result of two vectors' multiplication can be written as:

$$(v, x)(u, x) = \sum_{i=1}^{n} v_i x_i \sum_{j=1}^{n} u_j x_j = \sum_{i,j=1}^{n} v_i u_j x_i x_j \tag{2}$$

One can see that the right part of the expression is a quadratic function. Let us associate this function with the linear one according to rule:

$$x_i x_j = \sum_{k=1}^{n} \gamma_{ijk} x_k$$

So, let us rewrite (2):

$$\sum_{i,j=1}^{n} v_i u_j x_i x_j = \sum_{i,j=1}^{n} v_j u_j \sum_{k=1}^{n} \gamma_{ijk} x_k =$$

$$= \sum_{k=1}^{n} \left(\sum_{i,j=1}^{n} v_i u_j \gamma_{ijk} \right) x_k$$

This function can be represented as (w, x) and the components of vector w can be defined using the components of initial vectors v and u as follows:

$$w_k = \sum_{i,j=1}^{n} v_i u_j \gamma_{ijk} \tag{3}$$

In other words, we describe the specific kind of vectors' multiplication. According to (1), (2):

$$(v, x)(u, x) = (w, x)$$

Let us call the rule (3) the multiplication table and γ_{ijk} - the structural constants.

Such a determination of multiplication table is similar to the definition of algebra. But there is an important difference: the structural constants have no constraints such as commutativity, associativity and presence of "unit".

In order to avoid the evident question whether we can find the structural constants that satisfy (1) for every secret vector or not, let us indicate the method of its construction. Let us represent the structural constants as a set of vectors:

$$\gamma_{ij} = (\gamma_{ij1}, \ldots, \gamma_{ijn})$$

Thus, rewritten (1) looks as follows:

$$x_i x_j = (\gamma_{ij}, x) \ mod \ m, i, j = 1, \ldots, n \tag{4}$$

If we consider (4) as a set of linear equations with a given left part $x_i x_j$ and n^3 variables γ_{ijk}, these unknown variables are found ambiguous for every equation due to its non-trivial kernel.

The problem is the fact that in order to produce the real computations we need to disclose the structural constants. It is unobvious whether it is possible in this case to determine the secret vector with given constants. This question is equivalent to the question whether we can find a solution of the following system (if the coefficients γ_{ijk} are given):

$$x_i x_j = \sum_{k=1}^{n} \gamma_{ijk} x_k \ mod \ m \tag{5}$$

On the one hand, it is considered that solving the equation (2) in a finite field is a difficult task. But on the other hand, the system (5) consists of n^2 equations in reference to n variables, i.e. highly overdetermined. For highly overdetermined systems of equations it is expected that the solution is unique. There is one more argument to justify the complexity of the problem: the prime number is a secret, therefore, it is still unknown what modulus should be used in order to solve the system.

Thus, let us prove the following.

Theorem 3.1: The secret vector $x = (x_1, \dots, x_n)$ and the structural constants γ_{ijk} can be selected so that the system of equations (5) has at least n solutions.

In order to prove this theorem, let us give the construction of such a set of the structural constants. Let S be an arbitrary $n \times n$ matrix with the only constraint - it should be invertible by given modulus m. Then, choose two arbitrary columns of the matrix with i and j indexes (i and j may be the same). These columns match with two vectors - s_i and s_j respectively.

As it was mentioned, componentwise multiplication of the vectors $u = (u_1, \dots, u_n)$ and $v = (v_1, \dots, v_n)$ is defined by the following rule:

$$u \cdot v = (u_1 v_1, u_1 v_2, \dots, u_n v_n)$$

Let us define vector γ_{ij} as a solution of the equation

$$s_i \cdot s_j = S\gamma_{ij}$$

According to the invertibility of the matrix S, the solution can be rewritten:

$$\gamma_{ij} = S^{-1}\big(s_i \cdot s_j\big) \ mod \ m$$

All the columns of the matrix S satisfy the equation (5), where the structural constants are obtained as it is described above. As matrix S has n rows, we finally get n different solutions of the equation (5).

Remark 3.1: This construction is appropriate for any finite fields.

Remark 3.2: Since n different secret vectors correspond to the same set of structural constants, we can produce n secure computations simultaneously.

IV. Proposed Cryptosystem

In this section we consider the description of the proposed encryption that is based on modular arithmetic.

A. Basics

Firstly, let original message P be an integer number - we impose the only constraint: $P < M, M \sim 2^{64}$ in order to perform all the computations correctly. Then, let us define the encryption's secret key as a triple $(mods, \alpha, x)$, where

- $mods = (m_1, \dots, m_k)$ - a set of k moduli, m_i is prime $\forall i = 1, \dots, k$;
- $\alpha = (\alpha_1, \dots, \alpha_k)$ - a set of k arbitrary vectors needed for generating secret vectors x;
- $x = (x_1, \dots, x_k)$ - a set of k vectors with length n.

Thus, to encrypt P we should represent it as a set of residues $(P_1, \dots, P_k) : P_i = P \ mod \ m_i$ and after that construct a vector c_i for every P_i such that it satisfies the following condition:

$$(c_i, x_i) \ mod \ m_i = P_i$$

A set of vectors $C = (c_1, \dots, c_k)$ is a ciphertext for the initial number P.

B. Multiplication table

Before presenting the essence of the proposed encryption algorithm, let us describe the special multiplication table $T = (\gamma_{ijk})$ introduced in the previous section. Matrix T is used for the computations over ciphertexts in order to avoid the increase in the data lengths. We can work with the only multiplication table for all the moduli, but also we can generate k different tables for k different moduli. Let us consider this method for the chosen modulus m_i and fix the index i for all used terms; so then, we work simply with modulus m.

In order to generate such a table we need matrix S described in Section 2. Thus, computing the constants γ_{ijk} for every couple of i and j we get the specific multiplication table $T = (\gamma_{ijk})$ for the fixed modulus $m = m_i$.

Let us note one more feature of the multiplication table. If we construct matrix T as a non-symmetric matrix, we will get different results while computing $(a_i \cdot a_j) \cdot a_k$ and $a_j \cdot (a_i \cdot a_k)$. It means that the operation of multiplication has no associative and commutative properties. Also this fact invokes a non-deterministic character of the proposed encryption scheme.

C. Cryptosystem

Our fully homomorphic cryptosystem consists of three algorithms $(KeyGen, Enc, Dec)$, where

- $KeyGen$ - the probabilistic key generation algorithm that constructs the key;
- Enc - the encryption algorithm that takes initial message P, $mods$ - a part of the secret key and the multiplication tables T as the input parameters and returns a ciphertext C;
- Dec - the decryption algorithm that uses the secret key and the ciphertext C, returns the original message P.

1) Key Generation: As it was previously mentioned, the encryption key is secret and consists of the set of the relatively prime moduli and two sets of vectors. Let us consider the way of key generation in details.

Step 1. Let S be an arbitrary $n \times n$ matrix with non-zero determinant $det(S)$. Then we choose k relatively prime moduli $(m_1, \ldots, m_k)$ with a condition: $gcd(m, det(S)) = 1$. It is necessary in order to provide the invertibility of S by each modulus. Thus, matrix S for each modulus will be computed as follows:

$$S_i = (s_{ij}) \ mod \ m_i$$

Step 2. Then we should construct an arbitrary vector $\alpha = (\alpha_{i1}, \ldots, \alpha_{in})$ associated with modulus m_i using a rule:

$$\forall \alpha_{ij} \ \exists \alpha_{ij}^{-1} : \alpha_{ij} \cdot \alpha_{ij}^{-1} = 1 \ mod \ m_i$$

This rule means that every element of vector α_{ij} is invertible by chosen modulus m_i.

Besides we should provide the existence of at least two relatively prime elements in vector α_i in order to solve diophantine equations in the Enc algorithm.

A set of vectors α_i is also a part of the secret key.

Step 3. At the last step of key generation we compute x_i from the equation:

$$\alpha_i = Sx_i$$

Due to the fact that matrix S_i is invertible by modulus m_i:

$$x_i = (S_i^{-1}\alpha_i) \ mod \ m_i \ \forall m_i$$

Therefore, after key generation process we have k moduli $(m_1, \ldots, m_k)$ and the set of k secret vectors $x = (x_1, \ldots, x_n)$ constructed using the set of α_i. It is worth noting that the generation method is probabilistic due to the arbitrariness of S and α_i selection.

2) Encryption: The input parameters for this algorithm are the original message P - an integer number that satisfies the following constraint: $P < M, M \sim 2^{64}$, the secret key and the set of multiplication tables $(T_1, \ldots, T_k)$.

Step 1. Let us start with the computing the set $(P_1, \ldots, P_k)$ as follows:

$$P_i = P \ mod \ m_i \ \forall i = 1, \ldots, k$$

Step 2. Using vectors of the secret key $(\alpha_1, \ldots, \alpha_k)$, consider the equation:

$$P_i = (\alpha_i, y_i) = \alpha_{i1}y_{i1} + \cdots + \alpha_{in}y_{in} \tag{6}$$

Then compute the set of y_i as a result of the diophantine equation. Let us describe the way of solving such an equation in details. Due to the existence of two relatively prime components in every vector α_i the solution of this equation can be found as follows: let the position of two coprime integers be r and s, then choose random values for the coefficients $y_{iq} : q = 1, \ldots, n, q \neq r, q \neq s$ and substitute them into the equation (6). Thus, we get a linear diophantine equation with only two variables:

$$P_i - \sum_{q=1, q \neq r,s}^{n} \alpha_{iq}y_{iq} = \alpha_{ir}y_{ir} + \alpha_{is}y_{is} \tag{7}$$

The equation (7) can be solved, because the coefficients α_{ir} and α_{is} are relatively prime. Therefore, the values of the components y_{ir} and y_{is} can be computed using the Euclidean algorithm.

Also we can use the multiplication table T_i in order to solute such an equation. In this case we should only substitute x_ix_j in the formula (5) with P_i.

Step 3. Compute a cipher $C = (c_1, \ldots, c_k)$ using the following rule:

$$c_i = (y_i \cdot S_i) \ mod \ m_i$$

The result of the encryption algorithm is the ciphertext C that consists of k vectors of length n : $(c_1, \ldots, c_k)$. Thus, cipher C is a $k \times n$ matrix.

3) Decryption: The algorithm's input parameters are the ciphertext C, described previously, and the secret key.

Step 1. Compute a set $(P_1, \ldots, P_k)$ as follows:

$$P_i = (c_i, x_i) \ mod \ m_i \tag{8}$$

Let us prove the correctness of the equation (8) using previously given formulas of the encryption algorithm and the properties of the standard linear algebra:

$$(c_i, x_i) \ mod \ m_i = (y_iS_i, x_i) \ mod \ m =$$

$$= (y_i, \alpha_i) \ mod \ m = P_i$$

Step 2. As we have the set of P_i, apply the Chinese remainder theorem [3] and get the original integer number P that satisfies the next condition:

$$P \equiv \begin{cases} P_1 (mod \ m_1) \\ \vdots \\ P_k (mod \ m_k) \end{cases}$$

Let us consider the modification of the algorithm that provides the probabilistic character of the encryption in order to improve its security. Let C be a ciphertext for the initial number P. First of all, we compute a ciphertext corresponding to zero - C_0, then multiply it by an arbitrary coefficient θ. After that we add the result $\theta \cdot C_0$ to the ciphertext C:

$$C' = C + \theta \cdot C_0$$

Then C' is called a new ciphertext for the number P. As our encryption is fully homomorphic we may be sure the ciphertext C' is appropriate for P. So, to get the original message P, we should decrypt C' only. Thus, the proposed modification improves the complexity of the encryption algorithm. Such a modification is considered as a primary encryption algorithm. Its security evaluation will be discussed in the following section.

To conclude, in this section the details of the proposed fully homomorphic scheme were given. Briefly, let us mention the main features of this scheme again. The secret key is a triple $(mods, \alpha, x)$. We decided to perform all of the secure computations using modular arithmetic in order to avoid growth of the integers' size. Also the specific kind of vectors'

multiplication that allows performing arithmetical operations over ciphertexts without the increase in the resulting vectors' length was proposed. Then the probabilistic modification of our FHE scheme was described.

V. Encryption Security Evaluation

In order to analyze the complexity of the proposed FHE scheme, we provide some information about its efficiency:

- $O(k \cdot n^2)$ is the complexity of key generation algorithm;
- $O(n^3)$ is the complexity of multiplication table generation process;
- $O(n^2)$ is encryption algorithm's complexity;
- $O(k^2 \cdot n)$ is decryption algorithm's complexity.

In previous section it was mentioned that we might consider the proposed fully homomorphic encryption as a probabilistic one. The probabilistic encryption algorithm means that we get different ciphertexts if we encrypt the same plaintext more than once. Obviously such a modification prevents our scheme from common attacks, i.e. chosen ciphertext or plaintext attacks.

VI. FHE Applications

The proposed homomorphic encryption can be used in a multiple applications due to its practical allowance and acceptable data overhead. It's main purpose - as it was stated previously - to perform mathematical operations over encrypted data in untrusted and non-interactive environments without access to the encryption keys or initial data. So, the proposed solution can be practically used in the following cases of the secure computations.

A. Computation in Database

Databases and cloud databases, as a special case of cloud services, are affected by the same problem of keeping data confidentiality. Such a problem arises when a customer does not trust a database provider and/or an administrator or is not sure about security of connection between end user machine and database server [13]. Analogically, Fully Homomorphic and Order Preserving encryptions (OPE) can be applied to solve problem of keeping confidentiality of database entries. Properties of FHE and OPE allow users to perform any kind of computations (of course, with corresponding limitations) inside DBMS and the end user should decrypt only the result of selected data. Such an approach was implemented in MIT CryptoDB [14] and was positively acclaimed by the academy and the industry.

Alternatively we designed and developed a solution for secure Database [15]. We use proprietary developed OPE [16], proposed in this article FHE and strong deterministic encryptions. Main idea of our approach to secure database is to intercept user SQL queries on a flexibly configurable proxy server, encrypt vulnerable user's data and change the syntax of queries according to encryption's output ciphertext. Responses from DBMS are decrypted in a proper way and displayed to the user. The feature of granular security allows different encryptions to be applied to different columns in SQL table and perfectly accommodates user's requirements. Combination of implemented encryptions with carefully designed secure database architecture allowed us to achieve significantly low overhead of data flow and SQL queries' execution time. Estimated average overhead is around 20%.

This project allowed us to validate developed homomorphic encryption and to show its practical acceptance. Thus, we can perform secure computations over ecrypted data directly in protected database due to the properties of FHE. That is why such an application is primary for the proposed fully homomorphic encryption.

B. Cloud Computation

Cloud technologies are very popular and wide spread nowadays. Although customers of cloud services are very excited by cloud features and benefits that cloud has brought to enterprises, they are very concerned about security, particularity confidentiality, of data stored and processed in a cloud [7]. Those concerns are caused by several security issues of cloud technology in common, such as insider threat [8], possible security breach [9], intervention of special services into citizens privacy [10] and any other case of unauthorized access to vulnerable user data. There are multiple solutions [11][12] to described problem and one of them is usage of encryptions. Using homomorphic encryption or order preserving encryptions will allow business users to perform variety of operations over data stored in cloud data centers without necessity of massive computations on customers' side. Such a scenario will possibly lower expenses, while ensuring confidentiality of customer's data.

C. Constructing Public-Key Cryptosystem

Firstly we consider the application of fully homomorphic encryption for constructing linear and polynomial public-key cryptosystems. It is worth to note that we use the simplified method of the proposed encryption with fixed parameters: $k = 1, n = 4$. It means that we have the only modulus m and the only secret vector x.

The linear one is based on the Hill cipher [20]. In common way Hill cipher matches an original vector p to a ciphertext c according to the rule: $c = A \cdot p \ mod \ m$, where a square matrix A and a modulus m are secret. Besides, the matrix A should be invertible by the modulus m in order to provide the correctness of the decryption process. It is obvious that such a method is vulnerable to the plaintext attack. That is why the main idea of our approach is to hide the secret matrix A using the proposed FHE for its encryption. Also we encrypt the initial message with fully homomorphic algorithm E. Then we get a ciphertext, a result of public-key encryption, according to the rule: $c = E(A) \cdot E(p) \ mod \ m$.

The second, polynomial, cryptosystem is based on the analogue of the well-known RSA algorithm [19] where the modulus m is secret. Unfortunately this construction is unstable, but we can modify it using our fully homomorphic encryption. Thus, we propose to encrypt original number with the FHE algorithm E and after that raise the result of encryption to the power: $\big(E(p)\big)^e \ mod \ m$.

Let us consider the details of the polinomial cryptosystem via some examples.

1) Keys generation: Secret key consists of the components of the proposed homomorphic encryption's key:

$$m = 659$$

$$x = (176\ 657\ 361\ 197)$$

Public key includes an integer number $e = 3$ that is invertible by modulus $\phi(e)$ (where $\phi(a)$ is the Euler function for a) and the multiplication table γ_{ijk} that contains 16 vectors (or $4^3 = 64$ elements):

$$\gamma_{11} = (319\ 77\ 626\ 452)$$

$$\gamma_{12} = (80\ 182\ 161\ 229)$$

$$\gamma_{13} = (542\ 527\ 513\ 623)$$

$$\gamma_{14} = (2\ 148\ 241\ 557)$$

$$\gamma_{21} = (281\ 131\ 618\ 399)$$

$$\gamma_{22} = (568\ 414\ 276\ 590)$$

$$\gamma_{23} = (404\ 220\ 384\ 640)$$

$$\gamma_{24} = (238\ 252\ 389\ 179)$$

$$\gamma_{31} = (253\ 620\ 610\ 313)$$

$$\gamma_{32} = (304\ 88\ 55\ 421)$$

$$\gamma_{33} = (5\ 565\ 352\ 650)$$

$$\gamma_{34} = (63\ 390\ 604\ 279)$$

$$\gamma_{41} = (478\ 460\ 120\ 176)$$

$$\gamma_{42} = (78\ 568\ 258\ 224)$$

$$\gamma_{43} = (59\ 332\ 90\ 33)$$

$$\gamma_{44} = (432\ 103\ 198\ 222)$$

The size of secret and public keys is 2.5 Kb for the chosen parameters k and n.

2) Encryption: The initial number is an integer $p = 123$. The first step of the algorithm is to encrypt p using our fully homomorphic encryption. In other words, we should match p with a vector c that satisfies the following condition: $(c, x)\ mod\ m = p$.

$$123 \xrightarrow{Hom} \begin{pmatrix} 27458280 \\ 16546176 \\ 35555955 \\ 21767475 \end{pmatrix}$$

The second step is to raise the result of the FH encryption to the appropriate power e:

$$z = \begin{pmatrix} 27458280 \\ 16546176 \\ 35555955 \\ 21767475 \end{pmatrix}^3 =$$

$$= \begin{pmatrix} 360897386526156024805067154756 \\ 477019133423387912922438809475 \\ 488782414123179226098993372132 \\ 522900667259504641607843920158 \end{pmatrix}$$

Vector z is a ciphertext for the initial number p.

3) Decryption: First, let us multiply the ciplertext z and the secret vector x. It is obvious that as a result we get the initial number p raised to the power e:

$$(z, x)\ mod\ m = (c^e, x)\ mod\ m = p^e\ mod\ m$$

According to the example:

$$\begin{pmatrix} 360897386526156024805067154756 \\ 477019133423387912922438809475 \\ 488782414123179226098993372132 \\ 522900667259504641607843920158 \end{pmatrix} \cdot \begin{pmatrix} 176 \\ 657 \\ 361 \\ 197 \end{pmatrix}^T\ mod\ 659 =$$

$$= 510$$

Then, let us raise the result of the previous operation to the power d, where $d = e^{-1}\ mod\ \phi(m)$:

$$\left(p^e\ mod\ m\right)^d = p^{ed}\ mod\ m = p$$

Next, substitute the real values:

$$d = 3^{-1}\ mod\ 658 = 439$$

$$510^{439}\ mod\ 659 = 123$$

Finally we get the initial number $p = 123$.

Implementation of these cryptosystems demonstrates that all of the arithmetical calculations over encrypted data are correct. Also it proves that the multiplication of ciphertexts doesn't lead to the increase in dimension of multiplication results. This is the illustration of first practical use of the proposed FHE scheme.

D. Government Defensive Purpose

It is obvious that modern warfare needs a lot of computations. A part of these computations is done on machines using a software that are produced in foreign countries (for one fixed country), thus can not be fully trusted, because of possible hardware and software Trojans [17][18]. This problem of lack of trust can be solved by producing in a secure way the FH hardware encryptors. In the same time all untrusted computers will perform computations only over encrypted data.

All the mentioned applications are only examples of secure computation and described in this section as the illustrations of a wide area of the proposed homomorphic encryption usage.

REFERENCES

[1] C. Gentry, "A fully homomorphic encryption scheme," [Online]. Available: http://crypto.stanford.edu/craig/craig-thesis.pdf.

[2] C. Gentry and S. Halevi, "Implementing Gentry's Fully-Homomorphic Encryption Scheme," *in Advances in Cryptology - EUROCRYPT 2011,* pp. 129–148. DOI: 10.1007/978-3-642-20465-4_9. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-20465-4_9

[3] D. Knuth, *The Art of Computer Programming Seminumerical Algorithms,* vol. 2, Addison-Wesley Pub. Co., 1981.

[4] "Programming Computation on Encrypted Data," Broad Agency Announcement DARPA-BAA-10-81, Defense Advanced Research Projects Agency, 2010.

[5] R. Rivest, L. Adleman and M. Dertouzos, "On data banks and privacy homomorphisms," *in Foundations of Secure Computation,* 1978, pp. 169–180.

[6] D. Stehle and R. Steinfeld, "Faster Fully Homomorphic Encryption," *on Asiacrypt conference,* http://eprint.iacr.org/2010/299.pdf, 2010.

[7] "Cloud Computing Top Threats in 2013," *The Notorious Nine,* Cloud Security Alliance, [Online]. Available: https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf.

[8] W. R. Claycomb and A. Nicoll, "Insider Threats to Cloud Computing: Directions for New Research Challenges," *in Proceedings of the 2012 IEEE 36th Annual Computer Software and Applications Conference,* 2012, pp. 387–394. DOI: 10.1109/COMPSAC.2012.113. [Online]. Available: http://dx.doi.org/10.1109/COMPSAC.2012.113

[9] "Chronology of data breaches," Privacy Rights Clearinghouse, [Online]. Available: http://www.privacyrights.org/data-breach.

[10] "Interview with Whistleblower Edward Snowden on Global Spying," *Der Spiegel,* 2013.

[11] J. Zhou, "On the security of cloud data storage and sharing," *in Proceedings of the 2nd international workshop on Security in cloud computing,* 2014, pp. 1–2. DOI: 10.1145/2600075.2600087. [Online]. Available: http://doi.acm.org/10.1145/2600075.2600087

[12] A. J. Feldman, W. P. Zeller, M. J. Freedman and E. W. Felten "SPORC: Group collaboration using untrusted cloud resources," *in Proceedings of the 9th Symposium on Operating Systems Design and Implementation,* Vancouver, Canada, 2010.

[13] "OpenSSL Heartbleed Vulnerability," *Cyber Security Bulletins,* Canada, 2014.

[14] S. Tu, M. F. Kaashoek, S. Madden and N. Zeldovich, "Processing Analytical Queries over Encrypted Data," *in Proceedings of the 39th International Conference on Very Large Data Bases (VLDB),* Trento, Italy, 2013, pp. 289–300. DOI: 10.14778/2535573.2488336. [Online]. Available: http://dx.doi.org/10.14778/2535573.2488336

[15] K. Shatilov, V. Boiko, S. Krendelev, D. Anisutina and A. Sumaneev, "Solution for Secure Private Data Storage in a Cloud," *in Proceedings of the Federated Conference on Computer Science and Information Systems,* 2014, pp. 885–889. DOI: 10.15439/2014F43. [Online]. Available: http://dx.doi.org/10.15439/2014F43

[16] M. Usoltseva, S. Krendelev and M. Yakovlev, "Order-preserving encryption schemes based on arithmetic coding and matrices," *in Proceedings of the Federated Conference on Computer Science and Information Systems,* 2014, pp. 891–899. DOI: 10.15439/2014F186. [Online]. Available: http://dx.doi.org/10.15439/2014F186

[17] M. Tehranipoor and F. Koushanfar, "A Survey of Hardware Trojan Taxonomy and Detection," *in IEEE Des. Test,* 2010, pp. 10–25. DOI: 10.1109/MDT.2010.7. [Online]. Available: http://dx.doi.org/10.1109/MDT.2010.7

[18] R. Lehtinen, D. Russell and G. T. Gantemi, "Computer Security Basics," *O'Reilly,* 2006.

[19] A. Shamir, "A Polynomial Time Algorithm for Breaking the Basic Merkle-Hellman Cryptosystem," *CRYPTO,* 1982, pp. 279–288.

[20] L. S. Hill, "Cryptography in an Algebraic Alphabet," *The American Mathematical Monthly,* vol. 36, 1929, pp. 306–312.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 133–139
DOI: 10.15439/2015F59
ACSIS, Vol. 6

An Architecture for Secure Web Resource with Outsourced Database

Kirill Shatilov[1], Sergey Krendelev[1], Diana Anisutina[1], Artem Sumaneev[1], and Evgeny Ogurtsov[2]

[1]Department of Information Technology, Novosibirsk State University, Novosibirsk, Russia
Email: {shatilov, krendelev, anisyutina, sumaneev}@ccfit.nsu.ru
[2]Plesk, Parallels
Email: eogurtsov@parallels.com

***Abstract*—Security of outsourced data is crucial for businesses. To protect and secure outsourced database, as a part of dynamic web resource while maintaining site's work, we propose a solution featuring following key ideas. Firstly, we suggest to encrypt database content granularly with order preserving and homomorphic encryptions in order to conduct operation inside unmodified DBMS. Secondly, proposed solution implies presence of trusted intermediate component, responsible for SQL query precessing and site hosting tasks. Our approach has been validated through the implementation of complete web resource infrastructure with encrypted database. While web resource, using currently most popular content management system - WordPress, was functioning in normal mode, content of DMBS was secured. In this paper basic ideas of creating secure web resource will be discussed, as long as practical aspects. In addition proposed solution analysis using different metrics will be provided.**

I. Introduction

ENTERPRISES of all sizes are dependent on Internet for business. While some require simple brochure style web sites, others have need for a sophisticated dynamic web resources with vast amounts of data being processed. One of the most common types of dynamic web site is the database driven type. Outsourcing a database or whole hosting infrastructure brings many benefits - it lowers costs, increases reliability and accessibility, enables scalability. However delegating important data to third parties has some drawbacks; privacy issue is one of the most severe of them.

As long as insider threat exists, content of database may be stolen: malicious administrators or hosting provider's staff may capture or leak data[1], adversaries may obtain illegal access to sensitive information[2][3].

In this paper we present an architecture for secure web resource with remote DBMS to address listed security issues. Main effort is put towards securing database's content, while maintaining normal functioning of site's Content Management System (CMS). The key ideas of proposed secure web resource can be described as following:

1) *encrypted* contents of untrusted DBMS. Used encryptions are order preserving, homomorphic, so that selected operations over data can be preformed *inside* DBMS.
2) CMS and hosting software is set in *trusted* zone, where site administrators enforce their own security policies.
3) *interception* and *processing* of SQL queries from CMS and DBMS responses on trusted intermediate component. Transformed queries are executed on DBMS server.
4) *creating environment*, that would support queries transformation and maintain confidentiality, integrity and availability of resource's data.

This research was performed in Novosibirsk State University under support of the Ministry of education and science of Russia (contract no. 02.G25.310054)

The next section of paper discusses related work. After it, in Section 3, we describe the general ideas of proposed secure web resource. Section 4 gives some insight into encryptions, which are used in secure SQL queries processing. Following Section 5 presents practical aspects of the proposed solution. Analysis and evaluated results are listed in Section 7. Finally, the last section of this paper exposes future development options.

II. Related Work

While business processes are being tightly converged with information technology security procedures[4], multiple solutions for securing outsourced content were presented in Industry and Academy to fulfill demand of preserving data's privacy. Main accent is made to secure remote relation database by various encryptions.

Transparent Data Encryption (TDE) solution from DBMS developers, such as Microsoft[5] and Oracle[6], offers encryption at file level. TDE solves the problem of protecting data at rest, encrypting databases both on the hard drive and consequently on backup media, but does not protect contents of DBMS from access of illegal parties from database interface (e.g. theft of database user's authentication information).

In order to avoid limitation of database's full encryption several solutions for conducting queries over encrypted data were proposed. MIT CryptDB[7] presents a solution with support of multi-purpose encryptions based on, so called, "onion" database structure. Other solutions[8] offer fast execution of queries and search over encrypted data. However most of solutions lack fully homomorphic encryption and impose limitations on SQL queries.

III. Basic Principles

This section describes key ideas and distinctive features of the proposed solution.

Let us consider a web resource that is backed up by the database. The web resource is viewed as *secure* if its content is available only for legal users, or, in other words, confidentiality, integrity and availability of all resource's data are maintained constantly.

We propose an architecture for secure web resource with outsourced database. There are four main components:

- DBMS server
- crypto environment
- site engine (CMS) and web server
- client application

There is a clear division for architecture's components - they are either trusted or untrusted. The set of systems that can provide secure storage and processing facilities for confidential data is considered as *trusted* components. *Untrusted* environment is a set of systems, where security policy is setted up and enforced by third parties, therefore it can not be deputed to handle confidential information. As such, an untrusted system is one whose compromising will not lead to data leakage[9].

A. DBMS

The only untrusted component in proposed solution's architecture is outsourced DBMS. This DBMS is unmodified and plays a significant role in web resource functioning. All data stored in such DBMS in any particular moment of time must be secured. Additionally, all traffic to database and database's inner logs must not contain any vulnerable information, because it is supposed that malefactor can capture or monitor them.

B. Crypto environment

The key component of architecture is Crypto environment. Its main purpose to process SQL queries and DBMS responses while encrypting vulnerable data. Crypto environment is widely discussed in our first work[10]. Since first publication it has been improved in multiple aspects. Crypto environment consists of the following parts:

1) Encryptions subsystem: this component is an expandable set of encryption libraries with specified interfaces. Used encryptions are discussed in Section IV.

2) Metafile management subsystem: all keys and encryptions' meta data (types, names, count of output columns, other auxiliary information) are stored in meta file. Meta file subsystem's main purpose is to fulfil requests for keys and meta information of authorized components and to add new records for newly created tables and columns. Another subsystem's task is to encrypt/decrypt meta file when saving/loading from hard disk drive (meta file is always encrypted on Non-volatile Storage Devices for security reasons).

3) SQL queries processor: this component's objective is to parse SQL queries, reveal encrypted columns, encrypt data from this columns and to reconstruct resulting queries resting upon an encryption properties and encryption-specific math.

4) DBMS responses processor: decrypting selections from database.

As all encryption procedures are performed inside Crypto environment, DBMS never gets access to encryption keys.

C. Site engine

Site engine and web server software remains unmodified. The only subject to change is default SQL schema, that has to be encryption aware. All SQL- and response transitions are performed by Crypto environment. For site engine or any other SQL-backed application inside Crypto environment it is indistinguishable whether they work with usual DBMS or encrypted one.

D. Client application

End-users access secure web resource through ordinary browser. Access to resource is limited by login/password combination. Two factor authentication is the best option for hardening security.

E. Mechanism

Mechanism of web resource functioning is analogous to usual one - web server stores, processes and delivers web pages generated by site engine, that uses information from external database. But SQL queries from site engine are processed by Crypto environment, content of DBMS is encrypted, all queries are performed over encrypted data. During initialisation phase all "CREATE" queries are reconstructed and original data schema is mapped to encrypted one; all mapping parameters (meta information and encryption keys) are stored inside meta file. Analysis and reconstruction of further queries is based on this mapping parameters. All information requested by site engine for web page generation is presented in decrypted form by Crypto environment.

F. Configuration

There are two possible configurations of secure web resource - centralised and distributed.

1) Centralised: Schema of centralised secure web resource architecture is shown in Fig. 1. This concept implies that there is an additional trusted server, where site engine, web server and Crypto environment are set. All end users connect to this server using web browser. This proxy server has to be maintained in trusted environment by the administrators of secured web resource and can be a possible bottleneck for site access in case of high loads, but, on the other hand, proxy server allows exploitation of secure resource from mobile devices.

2) Distributed: Distributed configuration means that all trusted components are set on end user's device. Thus so called heavy client consists of http client, web server, site engine and Crypto environment. In case of distributed architecture additional meta file synchronisation mechanism is required in order to maintain consistency of meta information across different clients.

Distributed secure web resource architecture is illustrated in Fig. 2.

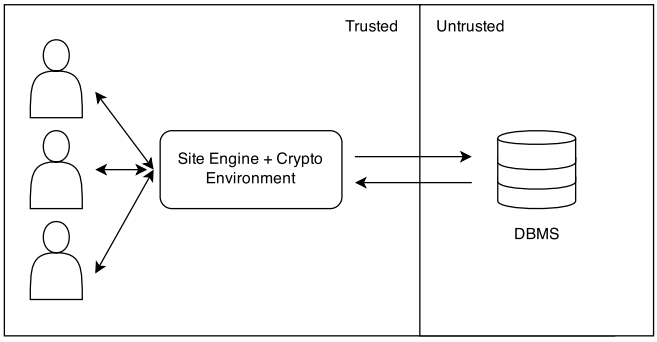


Fig. 1. Centralized Architecture

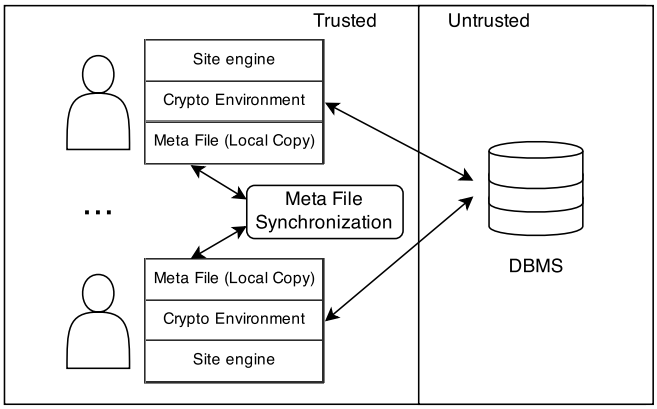


Fig. 2. Distributed Architecture

IV. Encryptions

A. Probabilistic

Probabilistic encryption is the use of randomness in an encryption algorithm, so that when encrypting the same message several times it will, in general, yield different ciphertexts. Data that are intended only for retrieval, but posses high value for users, can be encrypted in such way. Probable applications of such an encryption are comments' content or texts, where search operation is not essential. In the proposed solution, we use proprietary developed probabilistic block encryption. Implementation of Crypto environment is flexible, so that any other symmetric encryption can be used; for compatibility C++ interface-adaptor has to be implemented.

B. Deterministic

Deterministic encryption provides strong security, it leaks only which encrypted values correspond to the same data value. In site's data schema, deterministic encryption can be applied to user names and emails, search tags and password hashes. Usage of deterministic encryption for such data fields allows performing equality comparison and join operations, while it removes duplicate values from encrypted database (user names or emails must be unique, only one instance of each search tag exists).

C. Order Preserving

Order preserving encryption allows order relations between encrypted data items to be established, without revealing data itself. Such an encryption can be used for constructing secure indexes in database. With additional library functionality order preserving encryption can be applied to fixed-sized strings and dates with specified format.

Alternatively, order preserving encryptions can be probabilistic; encryptions with this property increase security, removing duplicate values from database, but also limiting operations held over data (equality comparison is not working properly). In proposed solution we use order preserving encryption discussed in [11].

D. Homomorphic

An encryption schema is called fully homomorphic, if it is able to evaluate an arbitrary function over ciphertexts. In this case decrypted value must match to a calculation result of the same function over plaintexts. The main feature of schema[12] that is used in the proposed solution is ability to define a strict upper bound of ciphertext size when performing calculations on it for both addition and multiplication. This fully homomorphic encryption is practically efficient and does not require huge computational or storage resources.

Homomorphic encryption can be applied to any integer data fields in site's SQL schema, that presents valuable information and supposed to be multiplied or added during web page generation (i.e. products' quantity and price in on-line shop).

V. Methods

This section features some examples of SQL queries handling to illustrate the challenge of correct processing of user queries with minimum limitations.

A. JOIN operation

In case, when some operations are intended to be performed over encrypted data across multiple columns and tables, following functionality can be used. In "CREATE" queries user must explicitly state that columns are members of one, so-called, "JOIN-group" in order to notify SQL queries processor to use one encryption key for all columns from group. The encryption must be deterministic for correct execution of further "SELECT" queries.

B. Probabilistic Order Preserving Encryption Handling

As it was previously stated, probabilistic order preserving encryption limits operation of equality comparison inside encrypted database. In order to avoid this limitation, following workaround was implemented.

Initial query example:

```
SELECT *
FROM table_name
WHERE column_name = value;
```

Output processed query:

```
SELECT *
FROM table_name
WHERE encrypte_column_name > OPE(value-1)
AND encrypted_column_name < OPE(value+1);
```

Here *OPE(value)* is considered as procedure of order preserving encryption.

C. Aggregate functions

If homomorphic encryption is used, some limitations are imposed. For example, aggregate function AVG (average value calculation) cannot be correctly performed inside DBMS. For correct execution of "SELECT" queries with average value calculation following transformation rules are applied.

Initial query example:

```
SELECT AVG(column_name)
FROM table_name;
```

Output processed query:

```
SELECT
SUM (encrypted_column_name),
COUNT (encrypted_column_name)
FROM table_name;
```

After response from DBMS is received, it consists of 2 values; first (sum) is decrypted and divided on second one(count), the result is passed to application.

In this section only a few examples of different SQL processing techniques were given to illustrate the mechanism of how Crypto environment work. However, some restrictions remain and it is a subject of future work to support all possible queries.

VI. Practical Approach

Practical implementation of the proposed concept in real application was essential for proving that such an approach for securing outsourced DBMS would work. In this section we discuss practical aspects of creating secure web resource with remote relational DBMS (MySQL[13]) using WordPress site engine[14].

WordPress is a free open-source blogging tool and content management system. Features include a plugin architecture and a template system. WordPress was used by more than 23.3% of the top 10 million websites as of January 2015[15]. WordPress is the most popular blogging system in use on the Web, at more than 60 million websites, according to official web site.

Proposed solution's practical implementation schema is illustrated in Fig. 3. Nginx[16] was used as a web server paired with a PHP FPM (FastCGI Process Manager); WordPress was configured to be used as a web page generation engine.

MySQL DBMS provides functionality of MYSQL Proxy, which is a software that intercepts traffic between database client (in case of web resource it is a php code of WordPress engine) and MySQL server(s) and can monitor, analyze or

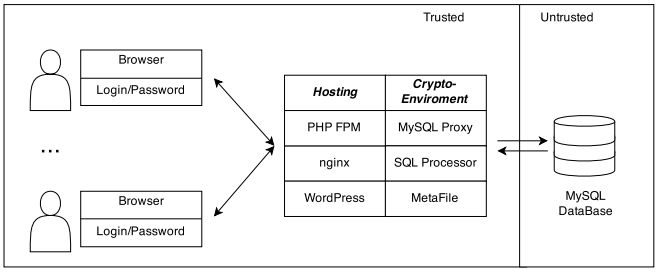


Fig. 3. Detailed schema of proposed solution

transform their communication. Its flexibility allows query analysis, query filtering and modification. Instance of MySQL Proxy was using .lua scripts, that called SQL queries and DBMS responses processors' procedures from prepared dynamic link library(.dll). All syntax analysis modules, encryption libraries were written on C++ programming language.

WordPress modifications were superficial and were made in three following directions:

A. Proxy

MySQl Proxy's IP address and TCP port were configured in WordPress as a default database, so that all queries generated by WordPress are processed by Proxy.

B. SQL schema modification

As it was stated in Section III, the application's SQL schema, which uses Crypto environment, must be encryption aware. Each column, which is intended to be encrypted, must be marked with encryption identification string in "CREATE" statement, as long as, JOIN-group name must be specified if needed:

```
CREATE TABLE wp_comments (
        --column name:
        comment_author
        --column type:
        tinytext
        --column constraints:
        NOT NULL
        --encryption id:
        encrypted_deterministic
        --join-group id:
        JOIN_GROUP wp_users_common,
... )
```

Encryptions applied to WordPress data schema are listed in Table I.

Homomorphic encryption was not applied to default WordPress configuration, but it is available for usage in various plugins with mathematical or financial functionality, e.g. for secure e-commerce resources.

Current way of encrypting WordPress SQL schema does not provide search over posts' content functionality, because probabilistic encryption is used for texts. But it is possible to add tags or keywords to post, so search will be performed over

TABLE I
ENCRYPTIONS, APPLIED TO WORDPRESS DATA SCHEMA

Field	Type	Encryption
tags, headers	text	deterministic
post, comments text	long text	probabilistic
post, comments, events date	date	OPE
user email, name	text	deterministic
user password	text	deterministic
ratings, order terms	integer	OPE

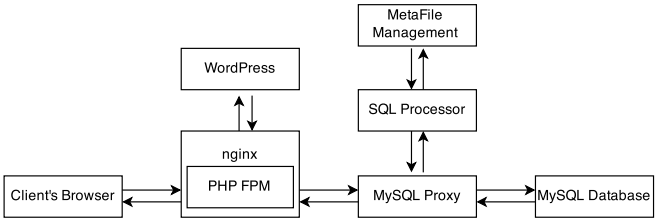


Fig. 4. Components Interaction

them. Mechanism of searching over encrypted data, proposed in [8], are similar - keywords are extracted from text and are indexed, search is held over set of keywords, not over the original text. So in this case, disabling search over texts is not a serious limitation.

Encrypting user names and passwords does not only protect database's login information's confidentiality, but also disables malicious database administrators from adding illegal users with right to access site's content.

C. Security Modifications

WordPress was configured so, that access to site was limited to users who are logged in or to users with an IP addresses from a specified set. In addition, search engines indexing was disabled.

To conclude, let us consider a typical components' interaction:

1) Client accesses web resource through browser using login/password.
2) Web server resolves http requests.
3) Web pages are generated by site engine using user's data from SQL database.
4) Site engine uses MySQL Proxy as a SQL database interface.
5) Scripts, used in Proxy, call SQL queries processor for syntax analysis and data encryption (or Selection processor to decrypt data selections).
6) Keys used for encryption or decryption are stored in meta file. Meta file management system resolves all requests by SQL queries processor and returns keys and meta information required for correct syntax processing.
7) After queries processed and all confidential data inside them are encrypted, proxy sends queries for execution inside DBMS.

Components interaction mechanism is shown in Fig. 4.

All web resource infrastructure was deployed on Microsoft Windows Server 2012 32bit OS. Current version of Crypto environment supports only Windows Operation System, while other components of secure web resource can be deployed across multiple platforms defined by MySQL and nginx vendors (e.g. CentOS, Ubuntu).

As a result of experiment, web resource was functioning, information uploading and retrieving was working correctly, while content of DBMS was encrypted. All queries, that was generated by WordPress, were compatible with Crypto environment and were executed correctly. However, slight time overhead was detected, and it will be discussed in following section.

VII. ANALYSIS

This section features analysis of the proposed solution and provides evaluated results.

A. Concept

There are 2 possible variants for configuring site infrastructure according to proposed solution - distributed and centralised. Heavy client, same as basic intermediate component, requires not more than 300 available hard disk space. Around 200 Mb of RAM is needed for normal functioning of heavy client. Centralised variant requires more RAM on machine with Web-server, site engine and Crypto environment in order to support sufficient quantity of simultaneous sessions, and optional high-availability component is recommended.

B. Performance

Developed solution uses encryptions and syntax processing, thus performance and data overhead exists.

1) Performance overhead: WordPress initialization with crypto environment took around 50% longer than unmodified WordPress. This significant increase is not critical, because initialization phase is performed very rarely. Content upload took ~20% more time, while content retrieval took less than 15% extra time.

2) Data overhead: Average database size increase is 60%. This high value of data overhead is caused by used encryptions.

C. Security

To prevent data leakage and strengthen security various mechanisms are available for the proposed solution. This part of a current section features overview of possible security breaches, evaluated damage to data confidentiality and suggested countermeasures.

Main purpose of suggested solution, as it was previously stated, is to protect vulnerable data in outsourced database. Adversary who gained access to database's content can not read private data, that are stored in database encrypted all of the time, but he can spoil or delete data. In addition he can calculate distribution of chosen cipher texts (applied for order

preserving encryptions) and perform evaluation of database size. Preventing successful cryptanalysis is possible by using strong encryptions.

Malefactor's access to Crypto environment is more serious security threat. Meta file is always stored encrypted on hard disk and is partially decrypted in RAM, thus crypto keys can be gained with low chances of success through RAM scanning. Accessing database proxy without database user login pass is ineffective.

Breaches on end users side are most severe. Malefactor can gain login and password of site users by exploiting various key loggers and relying on user's ignorance for security policies(password length, storing password in browser or in any other possibly insecure container). In this case full access to all sabotaged user's data is granted. After such a security breakthrough reencryption of meta file is the best option. This action will limit access of compromised user to none. To prevent this serious leakage scenario, several measures are recommended: responsible passwords management policy (constant passwords changing, reliable password length), regular reencryption of meta file, two-step authentication on client, increased security measures on users' workstations (anti viruses, proper firewall configuration, etc.).

To maintain integrity and availability of data in case of Meta file loss or corruption encrypted local meta file backups have to be done on user-defined schedule.

D. List of restrictions

Proposed solution solves problem of keeping confidentiality of outsourced data, while adding some constrains to normal functioning of web resource. In particular practical approach, using WordPress site engine following limitations are actual:

1) plugins usage: For each plugin that is working with confidential data special security patch is needed to secure plugin's data in DBMS.

2) backups usage: All backups must be performed in trusted zone, encrypted there and passed to DBMS. Otherwise an unencrypted backup in untrusted zone is an obvious threat to data confidentiality.

3) version dependency: WordPress data schema is changing through different versions, multiple tables and columns are being added or removed, thus any new confidential information, intended to stored in database, has to be encrypted; each version has to be revised and special security patch has to be applied.

VIII. Future Work

We presented our solution for secure site architecture, backed by database with minimal trust, that is immune to database theft and malicious insiders on DBMS side. Our research is aimed to delivery of a product, which will solve actual problems of information security, with minimal changes to site's hosting software and hardware. To achieve determined goal many changes must be done to the project in current state, including the following. First is automated or user defined selection of encryption for confidential data from any arbitrary SQL data schema. This improvement will discard some of described restrictions, while allowing us to present universal solution for multiple site engines and any custom plugins. Secondly, further security improvements to prevent any attack scenarios are going to be added to solution's features. Another direction for development is adopting client software for mobile platforms.

References

[1] W. R. Claycomb and A. Nicoll, "Insider threats to cloud computing: Directions for new research challenges," in *Proceedings of the 2012 IEEE 36th Annual Computer Software and Applications Conference*, ser. COMPSAC '12, Washington, DC, USA: IEEE Computer Society, 2012, pp. 387–394, ISBN: 978-0-7695-4736-7. DOI: 10.1109/COMPSAC.2012.113. [Online]. Available: http://dx.doi.org/10.1109/COMPSAC.2012.113.

[2] (2015). Data loss statistics, [Online]. Available: http://datalossdb.org/statistics (visited on 04/20/2015).

[3] (2015). Privacy rights clearinghouse. chronology of data breaches, [Online]. Available: http://www.privacyrights.org/data-breach/ (visited on 04/20/2015).

[4] G. Wangen and E. A. Snekkenes, "A comparison between business process management and information security management," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha L. Maciaszek, Ed., ser. Annals of Computer Science and Information Systems, vol. 2, IEEE, 2014, pages 901–910. DOI: 10.15439/2014F77. [Online]. Available: http://dx.doi.org/10.15439/2014F77.

[5] (2015). Microsoft. Transparent Data Encryption, [Online]. Available: https://msdn.microsoft.com/en-us/library/bb934049.aspx (visited on 04/20/2015).

[6] (2015). Oracle. Transparent Data Encryption, [Online]. Available: http://www.oracle.com/technetwork/database/options/advanced-security/index-099011.html (visited on 04/20/2015).

[7] R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan, "Cryptdb: Protecting confidentiality with encrypted query processing," in *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, ser. SOSP '11, Cascais, Portugal: ACM, 2011, pp. 85–100, ISBN: 978-1-4503-0977-6. DOI: 10.1145/2043556.2043566. [Online]. Available: http://doi.acm.org/10.1145/2043556.2043566.

[8] M. Sharma, A. Chaudhary, and S. Kumar, "Query processing performance and searching over encrypted data by using an efficient algorithm," *CoRR*, vol. abs/1308.4687, 2013. [Online]. Available: http://arxiv.org/abs/1308.4687.

[9] "The trusted systems problem: Security envelopes, statistical threat analysis, and the presumption of innocence," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 80–83, 2005.

[10] K. Shatilov, V. Boiko, S. Krendelev, D. Anisutina, and A. Sumaneev, "Solution for secure private data storage in a cloud," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha L. Maciaszek, Ed., ser. Annals of Computer Science and Information Systems, vol. 2, IEEE, 2014, pages 885–889. DOI: 10.15439/2014F43. [Online]. Available: http://dx.doi.org/10.15439/2014F43.

[11] M. Usoltseva, S. Krendelev, and M. Yakovlev, "Order-preserving encryption schemes based on arithmetic coding and matrices," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha L. Maciaszek, Ed., ser. Annals of Computer Science and Information Systems, vol. 2, IEEE, 2014, pages 891–899. DOI: 10.15439/2014F186. [Online]. Available: http://dx.doi.org/10.15439/2014F186.

[12] A. Zhirov, O. Zhirova, and S. F. Krendelev, "Practical fully homomorphic encryption over polynomial quotient rings," in *Internet Security (WorldCIS), 2013 World Congress on*, IEEE, 2013, pp. 70–75.

[13] (2015). MySQL official site, [Online]. Available: https://www.mysql.com/ (visited on 04/20/2015).

[14] (2015). WordPress official site, [Online]. Available: https://wordpress.org/ (visited on 04/20/2015).

[15] (2015). Usage statistics and market share of content management systems for websites, [Online]. Available: http://w3techs.com/technologies/overview/content_management/all/ (visited on 04/20/2015).

[16] (2015). nginx official site, [Online]. Available: http://nginx.org/ (visited on 04/20/2015).

Frontiers in Network Applications, Network Systems and Web Services

SYMPOSIUM SoFAST-WS focuses on modern challenges and solutions in network systems, applications and service computing. The Symposium builds upon the success of Frontiers in Network Applications and Network Systems (FINANS'2012) and 4th International Symposium on Web Services (WSS' 2012) held in 2012 in Wroclaw, Poland. These two events are now integrated into one event to fully exploit the synergy of topics and cooperation of research groups.

The topics discussed during the symposium include different aspects of network systems, applications and service computing. The primary objective of the symposium is to bring together researchers and practitioners analyzing, developing and administering network systems, with particular emphasis on Internet systems. Authors are invited to submit their papers in English, presenting the results of original research or innovative practical applications in the field.

Topics

Topics include (but are not limited to):

-
- Architecture, scalability and security of Open API solutions,
- Technical and social aspects of Open API and open data,
- Service delivery platforms - architecture and applications,
- Telecommunication operators API exposition in Telco 2.0 model,
- The applications of intelligent techniques in network systems,
- Mobile applications,
- Network-based computing systems,
- Network and mobile GIS platforms and applications,
- Computer forensic,
- Network security,
- Anomaly and intrusion detection,
- Traffic classification algorithms and techniques,
- Network traffic engineering,
- High-speed network traffic processing,
- Heterogeneous cellular networks,
- Wireless communications,
- Security issues in Cloud Computing,
- Network aspects of Cloud Computing,
- Control of networks,
- Standards for Web services,
- Semantic Web services,
- Context-aware Web services,
- Composition approaches for Web services,
- Security of Web services,
- Software agents for Web services composition,
- Supporting SWS Deployment,
- Architectures for SWS Deployment,
- Applications of SWS to E-business and E-government,
- Supporting Enterprise Application Integration with SWS,
- SWS Conversational Protocols and Choreography,
- Ontologies and Languages for Service Description,
- Ontologies and Languages for Process Modeling,
- Foundations of Reasoning about Services and/or Processes,
- Composition of Semantic Web Services,
- Innovative network applications, systems and services.

Event Chairs

Furtak, Janusz, Military University of Technology, Poland

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Legierski, Jarosław, Orange Labs Poland, Poland

Luckner, Marcin, Warsaw University of Technology, Poland

Szmit, Maciej, Orange Labs Poland, Poland

Program Committee

Benslimane, Sidi Mohammed, University of Sidi Bel-Abbès, Algeria

Chojnacki, Andrzej, Military University of Technology, Poland

Cocucci, Osvaldo, Orange Labs Products & Services, France

Fernández, Alberto, Universidad Rey Juan Carlos, Spain

García-Domínguez, Antonio, University of York, United Kingdom

Gibert, Philippe, Orange Labs Products and Services, France

Kaczmarski, Krzysztof, Warsaw University of Technology, Poland

Katakis, Ioannis, National and Kapodistrian University of Athens, Greece

Kiedrowicz, Maciej, Military University of Technology, Poland

Korbel, Piotr, Lodz University of Technology, Poland

Kowalczyk, Emil, Orange Labs, Poland

Kowalski, Andrzej, Orange Labs, Poland

López Nores, Martín, University of Vigo, Spain

Maamar, Zakaria, Zayed University, United Arab Emirates

Macukow, Bohdan, Warsaw University of Technology, Poland

Misztal, Michal, Military University of Technology, Poland

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 143–150

DOI: 10.15439/2015F91
ACSIS, Vol. 6

Analysis of video delay in Internet TV service over adaptive HTTP streaming

Marek Dąbrowski, Robert Kołodyński, Wojciech Zieliński
Orange Polska, Centrum Badawczo-Rozwojowe,
ul. Obrzeżna 7, 02-691 Warszawa
Email: marek.dabrowski@orange.com,
Email: robert.kolodynski@orange.com,
Email: wojciech.zielinski@orange.com

***Abstract*—The paper deals with OTT TV service based on adaptive HTTP streaming to deliver video content to various devices over unmanaged, best-effort IP network. One of major drawbacks of adaptive streaming technology is a significant video latency comparing to traditional TV broadcast. In this paper, causes of video latency in Internet TV architecture are identified and quantified by theoretical analysis of protocol behaviour and by testbed measurements.**

I. Introduction

VIDEO entertainment over the Internet has become a very popular service all over the world. Customers benefit from services offered by multitude of providers, which include pure OTT (Over The Top) players: local providers or worldwide giants like Netflix or Amazon, customer equipment manufacturers (Apple), content providers and TV stations (HBO, BBC), as well as legacy network and cable operators (Orange, Telefonica, Comcast). A VOD (Video on Demand) service is mainly offered by OTT service providers, but live TV is also gaining popularity, especially for watching live sports in the case of globally popular events. Television audience during events like 2014 Brazil Football World Cup is reaching its peaks all over the world and important part of this audience is watching on their PCs, smartphones and tablets [2].

Most commercially offered TV services over the Internet use so-called adaptive HTTP streaming technology [4]. It assumes that player is able to adapt to temporary network conditions by choosing among several profiles (versions of a stream encoded with certain bitrate), available on the server. The continuous stream is divided into fragments of certain size ("chunks") and delivered to clients using standard HTTP protocol. The format of delivered video fragments and manifest file (an index which allows clients to reach specific stream version) is governed by a streaming protocol, among which the most popular ones are: Microsoft SmoothStreaming, Apple HTTP Live Streaming, MPEG-DASH [4].

Work carried out within EUREKA CELTIC project NOTTS (Next generation Over-The-Top multimedia Services) [1]

A. Video latency in OTT TV

Live OTT TV service may suffer from significant video latency. A continuous video stream is divided into chunks (files), which suggests that certain amount of buffering must be applied in the streaming server. In addition, buffering is required in the end-device to circumvent network jitter and server overloads. As a consequence, end-to-end delay experienced by user is much larger than in the case of traditional broadcast, DTT, cable or IPTV service. Normally, a constant and stable delay is not a problem for viewers of movies or other non-live programs. However, the problem intensifies when someone is watching for example a football match, and may surprisingly hear his neighbors cheering over a scored goal, which he will only see on his screen in next minute, due to the delay introduced by OTT streaming. This problem may become more and more important with the advent of Social TV phenomenon. You may not really hear your neighbor shouting over the goal, but you will immediately see the comments posted by other viewers on Twitter, or you will see the news notification on your mobile phone, before you actually see the goal scored on the screen of your tablet. The issue of end-to-end (e2e) delay is thus becoming an important factor for overall QoE (Quality of Experience) of OTT services [3].

The discussed effect of e2e delay in OTT live TV is illustrated in Fig. 1, which depicts a photo taken while watching live transmission of a football match. The photo shows two screens at the same time: big TV screen connected to cable TV (DVB-C), and laptop screen, displaying an OTT TV service. One can see that match time shown on the laptop (OTT TV) is about 1 minute behind what is presented on the cable TV. Viewers of OTT TV are clearly experiencing a disadvantageous situation, not being able to follow the match truly live.

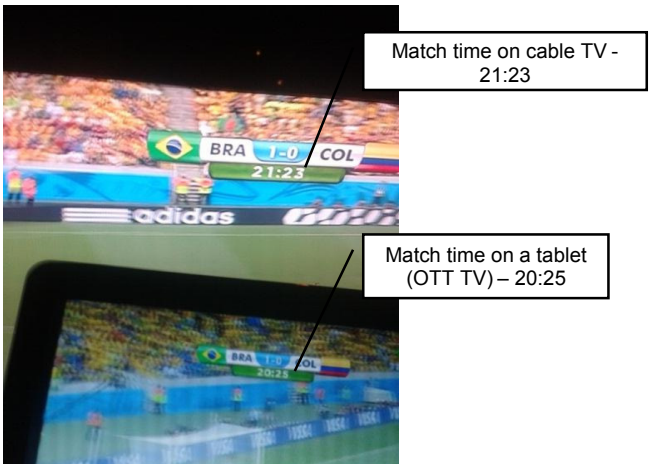


Fig. 1. Illustration of video latency problem in OTT Live TV

This study aims to identify, quantify and explain the causes of delay experienced by end user of OTT TV. The analysis will be supported by measurements in a testbed reproducing operational service architecture of OTT TV and video service offered by Orange Polska.

B. Delay budget in end to end delivery chain

Fig. 2 depicts typical architecture of OTT content delivery system and identifies major components which may contribute to e2e delay experienced by user.

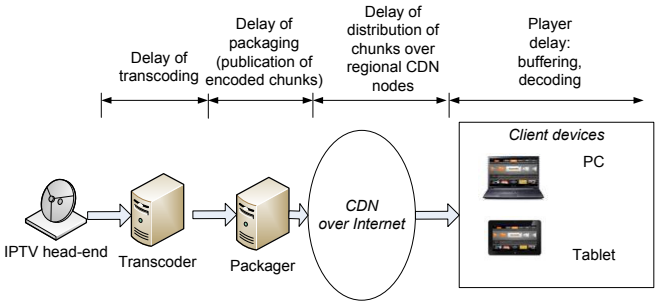


Fig. 2. Delay components in live OTT video

- **IPTV head-end**. Input content for OTT delivery chain is obtained from IPTV or satellite TV headend.
- **Transcoder** applies video compression, using several profiles appropriate for transmission over the Internet. H.264 is currently most popular compression standard, with HEVC (H.265) considered as future candidate.
- **Packager** applies streaming format (MS Smoothstreaming, MPEG-DASH, HLS,…). It divides continues stream to chunks of fixed size, prepares the manifest and publishes files on HTTP server.
- **CDN** (Content Delivery Network) is used for stream delivery to regional nodes in a wide area network. Since HTTP standard is used for message delivery, a typical Internet CDN is capable for supporting video streaming [6].
- **Video player** on the client device performs buffering, decoding and video playout. Length of receiving buffer, which is a major source of e2e delay, is a result of compromise between short e2e latency (small buffer), or better resilience against packet-level jitter and losses that may occur in the transport network (long buffer).

II. Adaptive Streaming Characteristics

In this section, essential characteristics of adaptive streaming technology will be analyzed from the point of view of impact on e2e video delay.

A. Transcoder behavior

The transcoder takes as input a continuous video stream, decodes it and encodes again, producing video fragments suitable for further processing by the packager. The encoding standard used in tested scenarios is H.264, the same as the input stream. The format of output file is fmp4, containing the amount of video equal to the packager's chunk duration. Remark that the encoder and packager use the same configuration of chunk duration, and are thus not totally independent in their operation.

Illustrative explanation of encoder impact on video delay is presented in Fig. 3. After the end of time period corresponding to chunk duration, the input video frames are stored in encoder's buffer. Next, they are processed by the encoder and the encoded chunk is saved on the encoder's storage disk as fmp4 file. The time of processing a video chunk by the encoder is non-negligible, and thus the video chunk that is saved on the disk at the output is delayed comparing to the input stream by the value of D_{enc}.

Remark that configuration of encoding profile may impact on the value of this delay, as better quality profiles surely require more processing at the encoder.

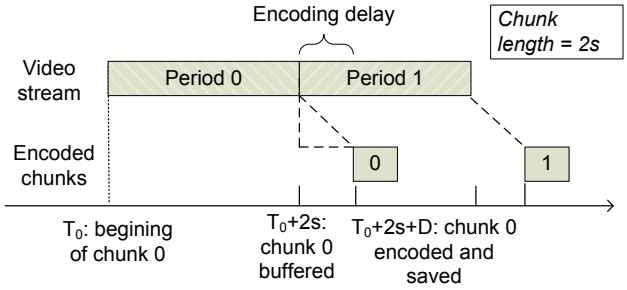


Fig. 3. Illustration of encoder behavior

B. Packager behavior

The following two parameters are crucial for operation of packager (see Fig. 4):

- **Chunk (fragment) length:** amount of video (expressed in time units) that is encoded and packaged in a single HTTP message transmitted over the network. The default value in Microsoft SmoothStreaming is 2s.
- Number of **lookahead** fragments: succeeding fragments that have to be collected by the packager before releasing a given chunk. The default value is 2.

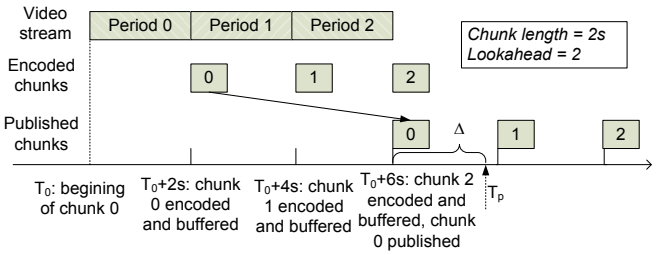


Fig. 4. Illustration of packager behavior

For the purpose of example, let us assume that the *chunk length* is 2s. The upper timeline in Fig. 4 shows a continuous video stream that is being served to the encoder. At the end of each period of 2s, the encoder produces a chunk (packet with encoded portion of the video). Thus, the chunk numbered 0, containing video period starting at *T0* and lasting 2s, is produced at time *T0+2s* and at the same moment it is stored by the packager in its internal buffer for further processing. However, since the *lookahead* parameter is set to 2, the packager will wait for next 2 consecutive chunks, because some information about these chunks must be built-in the header of chunk 0. Since chunk number 2 is available at time *T0+6s*, only then the chunk number 0 may be published and made available for clients.

Remark that player may request live stream at an arbitrary moment T_p (see Fig. 4). The first (newest) chunk available at this random moment T_p, is the chunk number 0, which is already aged *3*chunk length*, plus the duration of Δ, which is random. We may suppose that Δ is uniformly distributed between 0 and *chunk length*, with average value *chunk length/2*.

Thus, on average, the packager introduces delay equal to (l is the *lookahead*, and t_f is the *chunk length*):

$$D_{pack} = (l+1)\times t_f + \frac{t_f}{2} \tag{1}$$

C. Player behavior

Video player on the end-device is a major delay contributor in e2e delay budget. Microsoft SmoothStreaming introduces the following three parameters which have significant impact on behavior of the player when it starts receiving a live video stream:

- **Buffer**: size of receiver buffer (number of seconds of stored video). Default value is 5s.
- **Backoff**: when the player requests a live stream, it actually does not reach for the recent (current) video chunk, but rather for content that is delayed by a sum of backoff and offset parameters. Default value is 6s.
- **Offset**: together with backoff time, the value of this parameter determines playback delay in relation to actual "live" position. Default value is 7s.

When player requests to receive a live video stream, it downloads first a manifest file, which describes technical parameters necessary for the player to decode the stream and advertises the chunks that are available for download on the server. The timestamp of the latest (newest) chunk available within the manifest window is t_0.

However, the player does not normally reach for the chunk t_0. First, it goes back in time by the value of *backoff* plus *offset*. The sum of *backoff* and *offset* determines the timestamp of a chunk, from which the player starts downloading video fragments to fill its buffer (t_{start}). Now, the player immediately requests for next chunks, until it fills its buffer or reaches the limit determined by the offset value (player may not reach for chunks newer than "t_0 – *backoff*". We should now distinguish two situations: buffer ≤ offset, buffer > offset.

Player behavior when buffer is smaller or equal to offset

The player immediately requests for sufficient number of chunks to fill entire buffer. It gets them as fast as network bandwidth can support. Now it is ready to start video playback, beginning with the oldest chunk stored in the buffer. The timestamp of first chunk that will be displayed by player is:

$$t_{start} = t_0 - backoff - offset \tag{2}$$

The video delay as seen by the user will thus be t_0-t_{start}, that is:

$$D_{play1} = backoff + offset \tag{3}$$

Remark that chunk t_0 does not really contain "live" position of video stream, due to delay introduced previously by operation of encoder and packager.

Described behavior of the player is illustrated below in Fig. 6, which depicts chunks that are advertised when the player joins a live stream. The advertised window length is equal to 60s, which corresponds to 29 chunks of length 2s. The player parameters assumed for the purpose of example are: *buffer*=6s, *backoff*=6s, *offset*=20s.

The first (newest) chunk reached by the player has timestamp equal to t_0 – *26s*. Since the buffer size is smaller than the offset, all the buffer may be filled immediately by retrieving 3 chunks (6s) without waiting for any new chunks to be produced by the server. After retrieving enough chunks to fill the buffer to required length, the player starts playback, starting with the chunk t_{start}.

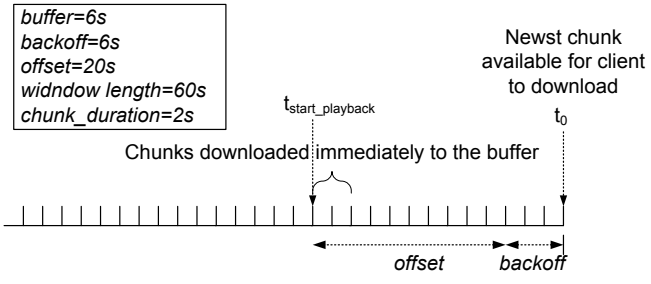


Fig. 5. Illustration of player behavior in the case *buffer* < *offset*

Player behavior when buffer is greater than offset

In the case when buffer > offset, the buffer cannot be immediately filled because the player is not allowed to fetch

chunks that are newer than t_0 - *backoff*. So, it immediately (as fast as the network bandwidth can support) fetches the amount of video chunks corresponding to the duration of an offset, and then waits for new chunks to arrive, to fill the remaining part of the buffer. After the time *(buffer – offset)* the buffer is filled and the player may start video playback, beginning from the chunk with timestamp t_{start}. But t_{start} is now additionally delayed from t_0 by *(buffer – offset)* because player had to wait that time to fill the buffer. So:

$$\begin{aligned} t_{start} &= t_0 - backoff - offset - (buffer - offset) \\ &= t_0 - backoff - buffer \end{aligned} \quad (4)$$

The video delay is equal to $t_0 - t_{start}$, that is:

$$D_{play2} = backoff + buffer \quad (5)$$

Described behavior of the player is illustrated below in Fig. 7. The advertised window length is equal to 60s, which corresponds to 29 chunks of length 2s. The example player parameters are the following: *buffer*=20s, *backoff*=6s, *offset*=7s.

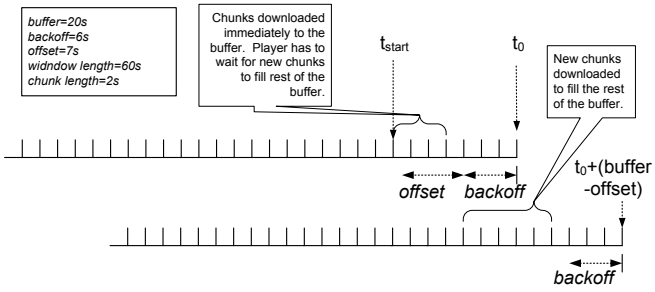


Fig. 6. Illustration of player behavior in the case *buffer* > *offset*

The first (newest) chunk reached by the player it the one with timestamp $t_{start} = t_0 - backoff - offset$. Since the buffer length is greater than the offset, all the buffer may not be filled immediately. The player thus retrieves rapidly (as fast as bandwidth may support) only "offset" portion of video chunks, and waits (*buffer-offset*) to gather enough newly arrived chunks to fill the rest of the buffer. Then, the player starts playback, starting with the chunk t_{start}.

Summarizing and merging equations (3) and (5) corresponding to different cases of player parameter settings, the formula for player delay can be written as:

$$D_{play} = backoff + \max(buffer, offset) \quad (6)$$

Remark that since packets sent over the network may be delayed or lost, causing a retransmission, the delay calculations should be treated as being "at least" values and the actual delay experienced may be greater than that.

Playback startup delay

We may expect that playback startup delay (between moment when user clicks on "play" button and moment when content actually starts playing) should grow with the size of the player buffer length. This is quite understandable because while joining the live stream the player must wait until the buffer is sufficiently filled, according to its configured value. More precisely, if the player buffer size is configured smaller that the value of offset, the player immediately asks for video chunks to fill the buffer completely. The chunks are thus downloaded almost instantly and the time player waits for filling the buffer is practically not observable. On the other hand, if the buffer size configured in the player is greater that the offset, the player cannot retrieve immediately the number of chunks required to fill the buffer. Thus it has to wait until sufficient number of new chunks appear on the origin server. The time it has to wait is equal to buffer minus offset (amount of video time that is missing in the current window stored on the origin server):

$$D_{play\_start} = \max(0, buffer - offset) \quad (7)$$

III. Experimental setup

A. Testbed architecture

A series of experiments have been performed for confirmation of protocol analysis from section II. Measurements have been carried out in a testbed, which reflects architecture of commercial OTT TV service of Orange Polska. Remark that names of equipment elements are only provided for information of the reader. It is not a goal of experiments described in this paper to evaluate particular vendor solutions. The characteristics that have been studied and measured are intrinsically related with generic technology (HTTP adaptive streaming) and only to a lesser extent depend on particular vendor implementation.

- Descrambler: Cisco DCM. It outputs a single decrypted TV channel in the form of IP multicast Transport Stream (TS), for further preparation of adaptive streaming content.
- Encoder: Ffmpeg v2.2 transcodes the content into H.264 stream packaged in fMP4 (fragmented MP4) format.
- Origin Server: Unified Streaming Platform (USP) v1.5.7 packages fMP4 content into the SmoothStreaming file format and produces manifest file. The content and manifest is served to clients by Apache HTTP server.
- CDN: Akamai Verivue.
- PC player: a web-based player developed in MS Silverlight.
- Mobile player: a reference application provided by the vendor of streaming player software.

B. End-to-end delay measurement

The instrumentation used for measurements of delay in OTT testbed is presented in Fig.8. The configuration of encoder machine allows us for adding current timestamp as an overlay, visually "burned" in video picture. This entry-point timestamp (measurement point A) can be visually compared with the current time on the user device (measurement point B), after passing entire delivery and

decoding process. Both clocks (in measurement point A and B) are synchronized with central clock by NTP protocol.

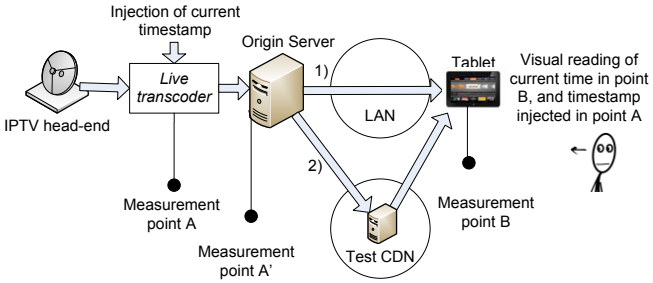


Fig. 7. End-to-end measurement testbed setup

The tesbed allows us for performing measurements including, or not, the impact of CDN. In the first case (path 1 on Fig.8) the end device retrieves content directly from the Origin Server, through a LAN laboratory network. The impact of network latency can thus be considered as negligible and CDN is totally eliminated. In the second case (path 2) the player reaches content through test CDN, consisting of a single cache node.

The test executor launches video player on a tablet connected to the test network and manually (visually) reads the measurement results.

Current timestamp in point A (T_A) is embedded in each video frame. Simultaneous readout of this timestamp and current timestamp (absolute time) in measurement point B (T_B) lets us estimate total time of processing given video frame in entire content delivery chain. The e2e delay can be calculated in any given moment as:

$$D_{e2e} = T_B - T_A \tag{8}$$

Accuracy of assumed measurement method is limited due to visual readout of timestamps. Normally, human tester may read the timestamp from computer and tablet screen with granularity of around 1 second. More fine-grained measurement of time would require some automation of the method and more precise instrumentation. For limiting the impact of human error, each measurement has been repeated several times. Taking into account that typically e2e delay in OTT delivery chain may be in the order of 20 sec – 1min, the granularity of assumed method seems to be sufficient.

Remark that presented method actually measures e2e delay, which is a sum of several delay components:

$$D_{e2e} = D_{enc} + D_{pack} + D_{CDN} + D_{play} \tag{9}$$

Additional actions must be taken to split it into particular components, as explained below.

C. Encoder delay measurement

Factors which impact on delay introduced by encoding process include: encoder implementation efficiency, performance of hardware on which it is being run, whether the encoder itself is software or hardware based, numerous parameters that can be set on the encoder and may alter its performance.

The transcoder installed in the testbed and used in the scope of this study is a software-based solution ffmpeg 2.2, running on Centos 6.5 64 bit system, installed as virtual machine (Oracle VM VirtualBox, 1GB RAM, 2 CPU). The virtual machine was running on Windows Server 2008 R2 Standard 64-bit (HP ML150: 2xIntel Xeon CPUE5504 2GHz, 4GB RAM).

Fig.9 gives more details into the configuration of transcoder, presenting video processing steps and the detailed points where timestamp was embedded into the video for purpose of measuring delays.

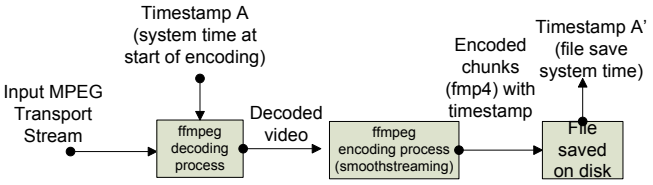


Fig. 8 Encoder configuration (with timestamp embedding) for measurements of encoding delay

As the first step within the transcoder module, FFmpeg decodes the input stream to produce raw video, which is then encoded to smoothstreaming compatible format by the encoder. However, prior to encoding, ffmpeg process "burns" a timestamp in each produced video frame. The timestamp value in point A (T_A) corresponds to current system time when given frame has entered the transcoding process.

The output of the encoder is an fmp4 file, containing amount of video corresponding to a duration of a chunk. Remark that although the encoding and packaging processes are logically separated, the encoder is not totally independent of the packager as it prepares an encoded portion of the video which suits the packager's chunk size.

The encoded chunks (in fmp4 format) are then saved to the storage of transcoder machine. The time of file modification is recorded in the file system as a normal operation of the computer's operating system and it is considered as a timestamp in point A' ($T_{A'}$). By comparing timestamp A' of a chunk, with timestamp A of the last video frame of each chunk, we can estimate the delay introduced by the whole transcoding process.

$$D_{enc} = T_{A'} - T_A \tag{10}$$

D. Packager delay measurement

In order to evaluate impact of packager in the e2e delay budget, we have performed a set of measurements using the same methodology as described for the e2e delay, but with a specific setting of player parameters. By setting *buffer=1*, *offset=0* and *backoff=0* we reduce the impact of player practically to zero. Without backoff and offset the player reaches for the newest available chunk while joining the live stream (see Fig.10). Since this single chunk is sufficient to

fill the buffer, the player may start playback immediately after receiving it. In a real network situation such configuration is not recommended since it is very susceptible to network impairments. However in the "idealized" testbed environment we were able to properly play a live stream with such non-realistic parameter setting.

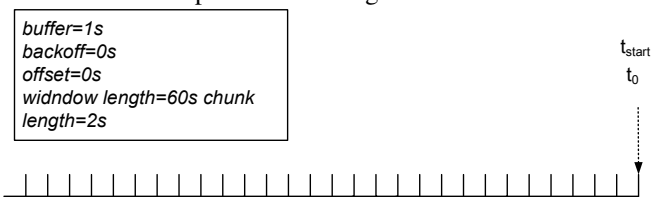


Fig. 9. Illustration of player behavior in the case buffer=1, offset=0, backoff=0

Since player starts playback immediately after receiving the newest chunk from the origin server, we may expect that observed delay in measurement point B is only related with the packager and encoding delay ($D_{play}=0$). So, the assumed procedure was to measure delay in e2e relation (without CDN: D_{CDN}=0) and substract from it known value of encoding delay, obtained by previous measurement of D_{enc} in the same setup.

$$D_{pack} = D_{e2e} - D_{enc} \tag{11}$$

E. CDN delay measurement

As depicted in Fig.8, testbed configuration allows for performing measurements with, or without CDN in the delivery chain. The assumed indirect methodology for evaluating impact of CDN itself assumes comparing the end-to-end delay results measured "with" and "without" CDN.

$$D_{CDN} = D_{e2ewithCDN} - D_{e2ewithoutCDN} \tag{12}$$

F. Player delay measurement

The methodology of evaluating impact of the player alone assumes performing measurements in end-to-end mode (see section III.B), without CDN ($D_{CDN}=0$) and substracting from result the values of delay of encoder and packager, known from prior measurements in the same setup. Thus,

$$D_{play} = D_{e2e} - D_{enc} - D_{pack} \tag{13}$$

IV. MEASUREMENT RESULTS

A. Encoder delay impact

The encoder delay has been measured according to the methodology described in section III.C, with chunk length changed from 1s to 10s (remark that although chunk length is a parameter of packager, the encoder must be configured accordingly in order to produce encoded video fragments that are suitable for the packager).

In addition, several encoding profiles have been tested (baseline, main), with several settings of ffmepg tool encoding parameters (medium, fast, ultrafast). The results of experiments are presented in Table I. Reported measured delay is an average calculated over five repetitions of each experiment.

TABLE I.
MEASURED ENCODER DELAY

Encoder profile	Chunk size	Avg measured delay [s]
Baseline, fast	1	1.49
	2	1.74
	5	1.78
	7	1.76
	10	1.79
Baseline, medium	1	1.54
	10	2.11
Main, fast	1	1.73
	10	1.94
Main, ultrafast	1	1.36
	10	1.19

The encoder delay in testbed environment is roughly between 1.5 and 2 seconds. We recognize that obtained results could differ for another encoder type, running in different environment. Therefore, we stress that the results are relevant for particular hardware/software configuration of our testbed and cannot be generalized in straight forward way to other types of encoders available on the market.

B. Packaging delay impact

The packager delay has been measured as described in section III.D, with various chunk length set on the packager (1 to 10s), and with different values of lookahead parameter (1, 2, 4, 6 fragments).

The results are presented in Table II and Fig.11 (subset of results with *lookahead*=2). The measured delay is compared with theoretical value D_{pack} from equation (1).

TABLE II.
MEASURED PACKAGER DELAY

Chunk length [s]	Lookahead	Measured packager delay [s]	Theoretical value of D_{pack} (eq.1) [s]
1	2	5.51	3.5
2	2	7.86	7
5	2	18.22	17.5
7	2	25.44	24.5
10	2	34.61	35
2	1	6.46	5
5	1	13.02	12.5
7	1	19.64	17.5
10	1	24.61	25
2	4	12.66	11
10	4	57.41	55
5	4	28.82	27.5
2	6	16.26	15

One can observe that measured packager delay can be quite well approximated by formula (1).

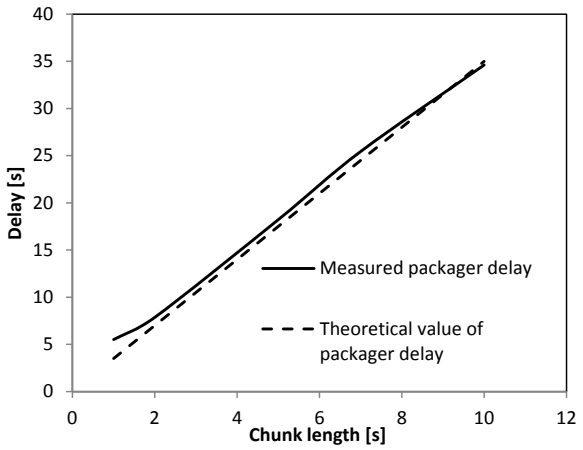


Fig. 10 Measured packager delay, with lookahead=2

C. CDN delay impact

The delay impact of CDN has been assessed using methodology described in section III.E. Measurements were done with different value of chunk size on packager (1, 2, 5, 7, 10s) and with fixed value of *lookahead* = 2. The results of measurements, averaged over 5 repetitions of each experiment, are presented in Table III. The average delay impact of CDN is indirectly estimated as difference of delay measured with, and without CDN in the testbed.

TABLE III.
MEASURED CDN DELAY

Chunk length [s]	Average delay with CDN [s]	Average delay w/out CDN [s]	Average delay of CDN [s]
1	6	6.6	**0.6**
2	8.8	9	**0.2**
5	14.4	15.8	**1.4**
7	18.6	22	**3.4**
10	26.2	30.8	**4.6**

We can observe that CDN does not introduce significant delay in the end-to-end chain. The observed delay is between 0.2 and 4.6s, depending on the chunk length.

D. Video player delay impact

Delay has been measured in end-to-end mode, without a CDN (see section III.F). The delay of encoder and packager has been eliminated by subtracting the results of previous measurements from section IV.A and IV.B, performed with the same parameter setting.

In first series of experiments, the delay was measured with different values of *buffer* length in video player. The values of two other player parameters were fixed to *backoff*=6s, *offset*=7s. Note that the values *buffer*=5s, *backoff*=6s and *offset*=7s are considered as default in the Microsoft Smoothstreaming protocol. Two values are reported as result of experiments (see Table IV and Fig.12):

- *"Start delay"* corresponds to stream startup time. It was measured with a stop watch, as time between clicking "play" on a player, and actual start of video playout. Reported value is an average over 5 repetitions of each experiment.
- *"Player delay"* corresponds to the observed difference between watched video and actual "live" position, measured as described in section III.F. The reported values are an average and minimum over 5 repetitions of each experiment.

TABLE IV.
MEASURED PLAYER DELAY AS FUNCTION OF BUFFER LENGTH

Player parameters			Startup delay [s]		Player delay [s]		
buffer [s]	backoff [s]	offset [s]	Avg	D_{play_start} (eq.7) [s]	Average	Min	Theoretical D_{play} (eq.6) [s]
3	6	7	1.26	0	14.06	13.26	13
5	6	7	1.75	0	14.26	13.26	13
7	6	7	2.56	0	15.46	15.26	13
10	6	7	3.67	3	18.26	15.26	16
13	6	7	4.10	6	17.46	16.26	19
15	6	7	6.68	8	19.26	18.26	21
17	6	7	7.73	10	20.26	19.26	23
18	6	7	10.81	11	23.46	23.26	24
20	6	7	11.73	13	25.66	24.26	26
25	6	7	15.59	18	28.46	27.26	31
30	6	7	23.00	23	36.46	36.26	36

One may observe that the player delay can be quite well predicted using formula (6).

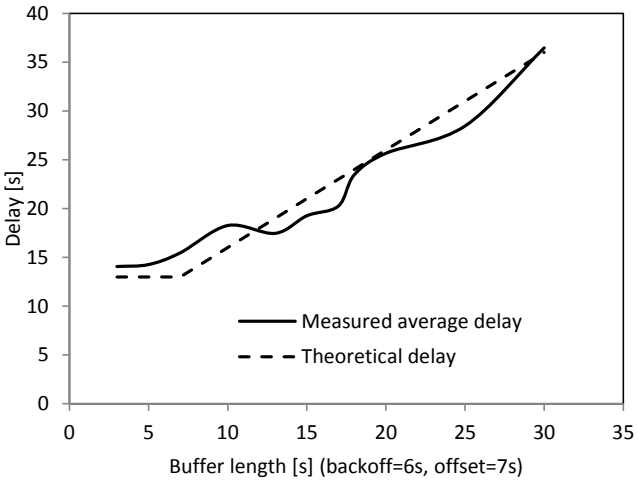


Fig. 11. Measured player delay as function of player buffer length

Fig. 13 shows the measured playback startup delay. We may observe that it is well approximated by formula (7).

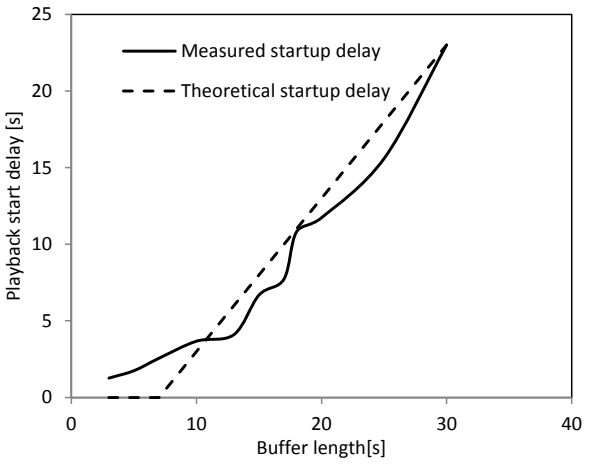


Fig. 12. Measured playback start delay as function of player buffer length

In the second set of experiments, the player *offset* parameter was varied, with fixed *buffer* length equal to 5s, and fixed *backoff* equal to 6s. The results are presented in Fig.14.

Note that the playback startup delay does not significantly depend on value of *offset* parameter, because in this particular case the buffer is usually smaller than the offset (except from the first two measurements).

Once again the results confirm validity of formula (6) for predicting latency of the player. Above the value of *offset*=60s chunks that should be retrieved are out of the range of advertised window, which means that player wants to download chunks that are too old and do not exists anymore on the server. Thus, formula (6) does not apply.

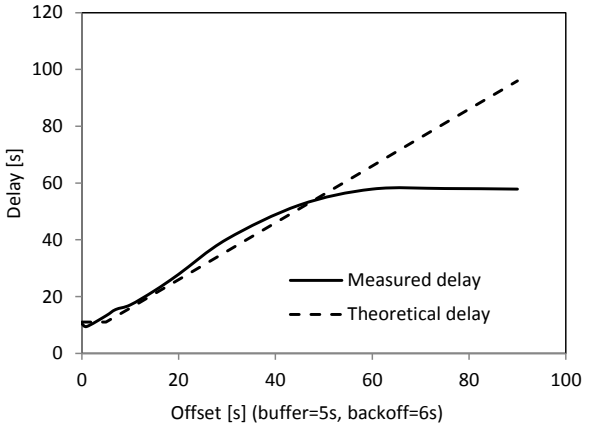


Fig. 13. Measured player delay as function of player buffer offset

In the last set of experiments, the player *backoff* time has been varied in the range from 0 to 90s, with constant *buffer*=5s and *offset*=7s. The results are presented in Fig.15.

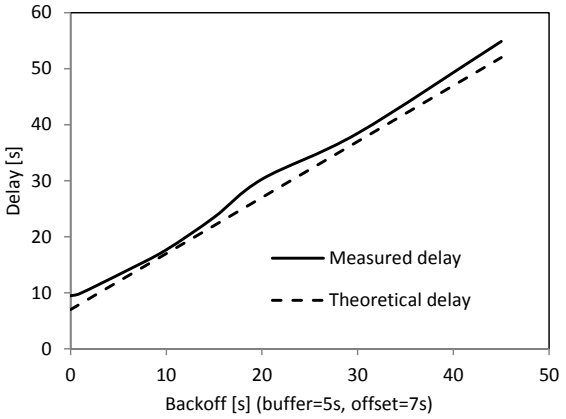


Fig. 14. Measured player delay as function of player buffer backoff

Once again, the results confirm that the delay introduced by the player can be estimated by equation (6).

V. CONCLUSIONS

The paper has presented results of analysis and measurements of user-perceived delay in Live TV service delivered over the Internet using adaptive HTTP streaming technique. The following elements of content delivery architecture have been identified as major contributors to overall latency: source stream transcoding, packaging (applying adaptive streaming format), delivery over CDN, and buffering on end-device.

Presented results are of analytical as well of experimental type and may have practical importance for video service providers as hints for setting key system parameters, taking into account both technical constraints and user Quality of Experience.

REFERENCES

[1] EUREKA / CELTIC NOTTS: http://projects.celticplus.eu/notts/

[2] Broadband TV news: http://www.broadbandtvnews.com/2014/06/12/world-cup-matches-to-set-new-streaming-records/, last accessed on 23.04.2015

[3] R.Merkuria, P.Cesar, D. Bulterman, "Digital TV: The Effect of Delay when Watching Football", EuroITV'12, 10th European Conference on Interactive TV and Video, Berlin, July 2012, http://dx.doi.org/10.1145/2325616.2325632

[4] T.Stockhammer, "Dynamic adaptive streaming over HTTP: standards and design principles, ACM MMSys '11, dx.doi.org/10.1145/1943552.1943572

[5] Microsoft SmoothStreaming: http://msdn.microsoft.com/en-us/library/microsoft.web.media.smoothstreaming.smoothstreamingmediaelement.liveplaybackoffset%28v=vs.95%29.aspx , last accessed on 26.06.2015

[6] K.Kaczmarski, M.Pilarski, "Content Delivery Network Monitoring", Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pages 633 – 639, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6354443&isnumber=6354297

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 151–158

DOI: 10.15439/2015F336
ACSIS, Vol. 6

Migration Towards Broadband PPDR Networks

Henryk Gierszal*, Piotr Tyczka†, Karina Pawlina†, Krzysztof Romanowski†, Damien Lavaux‡,
John Burns§, Val Jervis§, Luis Teixeira¶, and Andre Oliveira¶
*Adam Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland
Email: gierszal@amu.edu.pl
†ITTI Sp. z o.o., Rubież 46, 61-612 Poznań, Poland
Email: {piotr.tyczka, karina.pawlina, krzysztof.romanowski}@itti.com.pl
‡DSC Advanced Studies, Thales Communications & Security SAS, Gennevilliers, France
Email: damien.lavaux@thalesgroup.com
§Plum Consulting, 10 Fitzroy Square, London, United Kingdom
Email: {john.burns, val.jervis}@plumconsulting.co.uk
¶TEKEVER, Rua da Leziria 1, 2510-080 Obidos, Portugal
Email: {luis.teixeira, andre.oliveira}@tekever.com

***Abstract*—This paper presents a study on development and evolution of Public Protection and Disaster Relief (PPDR) communication networks in the perspective of the next 15-20 years. A need for modern, reliable mobile communications systems for PPDR offering a wide range of services, yet harmonized, is commonly recognized at both country and European level. There are therefore a number of activities in this area currently being undertaken. In this paper we discuss several technical aspects of migration of PPDR networks towards broadband systems and describe the most likely upgrade scenario based on an evaluation of migration costs.**

I. Introduction

PUBLIC Protection and Disaster Relief (PPDR) is the general designation given to a range of public safety services: medical emergency services, police squads, fire brigades, etc. Secure and reliable wireless communication between personnel and sections of PPDR agencies, as well as between different PPDR agencies is a vital element of their successful operation and ensuring safety of PPDR staff both in routine and emergency situations. Another crucial challenge is to make these communications suitable for "media type" services, that are increasingly utilized over public mobile networks, to improve the range of information available to the PPDR end-users.

Nowadays PPDR organizations are facing the challenge of migrating from narrow- and even wideband Public Safety Communication (PSC) systems to broadband ones because different public services or utilities increasingly need broadband solutions, e.g. to transmit video streams from an incident scene to the headquarters or benefit from applications of augmented reality. Simple replacement of current network elements to broadband ones is not feasible because broadband transmission needs wider bandwidths that can only be allocated in higher frequency band. Also the adoption of higher frequencies will reduce the maximum range of cells and require additional sites. In addition the broadband infrastructure will need to be upgraded so it can support the higher capacity required to support the higher bandwidth traffic and also the additional sites.

This work was supported by European Union in the framework of the FP7-SEC-2012.5.2-1 program – project "Public Protection and Disaster Relief – Transformation Center – PPDR-TC" (contract no. 313015).

In this paper, based on work performed under the EC FP7 project "Public Protection and Disaster Relief – Transformation Center – PPDR-TC", we present a study on technical, business and organizational aspects of development and evolution of PPDR mobile communication networks. It takes into account the likely developments of wired and wireless communications networks over the next 15-20 years as well as the currently available technologies.

The paper is organized as follows. In Section II we present the PPDR requirements for future networks that result from their needs and available technologies. Section III describes the architecture of the future PPDR systems that meets both end-users requirements and possible business models. In Section IV the migration path towards broadband PPDR networks is proposed and discussed. Section V gives an example of expenditures for such a migration. Finally, Section VI contains conclusions.

II. PPDR Requirements

Modern telecommunication networks have evolved as service-oriented infrastructure in order to meet end-users' requirements that in fact are related to different use-cases. Literature survey [1]-[3] and meetings with stakeholders led to the definition of eight typical communication scenarios:

- Communication Scenario A: between a central control station and field personnel at an incident location,
- Communication Scenario B: between PPDR vehicles and an incident location or control station,
- Communication Scenario C: between individuals at an incident location,
- Communication Scenario D: between different PPDR entities (e.g. police, fire, ambulance),
- Communication Scenario E: accessing information from the Internet or other external data sources (including

corporate intranets),
- Communication Scenario F: communication in enclosed spaces (e.g. tunnels or basements),
- Communication Scenario G: communication with remote locations (e.g. mountains),
- Communication Scenario H: communication with or between machines (e.g. remotely controlled vehicles).

The above listed scenarios require different traffic volumes to be supported depending on the operational circumstances in which the PPDR agencies are involved. Three categories of operational activities were defined by end-users:

- Routine day-to-day activities – traffic volume is low or moderate (no network congestion observed),
- Major events – higher communication needs have to be provided as compared to routine day-to-day activities but the location and requirements are known in advance, therefore some planning can be done earlier,
- Major incidents or disasters – higher communication needs have to be available at very short notice and the location and requirements are not known in advance.

The bearer services (i.e. basic services for transport of information payloads) that may be required by PPDR users was another type of classification considered in the PPDR-TC project. On the basis of the opinions provided by end-users and other analysis [3, 4], the following five categories were identified:

- Voice,
- Narrowband (NB) Data (e.g. for messaging),
- Broadband (BB) Data (e.g. for sending or receiving images or large files, and for accessing databases),
- Video (similar to broadband data but likely to be more demanding in terms of latency),
- Use of repeater stations to extend coverage, e.g. into enclosed or remote areas.

Broadband services are also needed for video and challenging services.

Finally, a series of functionalities that should be provided by broadband PPDR networks was identified and grouped into six sets as shown in Table I. Voice services will always be required irrespectively of the other types of services available for PPDR. The throughput threshold used to distinguish between narrow- and broadband transmission was set at 384 kbps, however in many definitions it has been already increased to 1 Mbps and recently even to 4 up to 25 Mbps [5]. Broadband transmission is required by video services and other new potentially challenging services. Repeater (relay) services are linked with communication scenarios F and G discussed above. All services should be available within the whole PPDR network of, at least, national coverage.

PPDR organizations can use services listed in Table II for many applications identified in [6]. Many of them need broadband transmission but there are still plenty of applications for which narrowband transmission is sufficient.

Table I
REQUIRED FUNCTIONALITIES FOR BROADBAND PPDR NETWORKS

Groups	Functionalities
Voice (common PPDR voice services)	push-to-talk
	private call
	group call
	emergency/priority call
	call retention/busy queuing
	direct mode operation
	ambience listening
	voice over the Public Switched Telephone Network (PTSN)
	area selection/Dynamic Group Number Assignment (DGNA)
Narrowband data (data transmission up to 384 kbps)	messaging and notifications
	low resolution photos
	automatic telemetrics
	location-based information
	mobile workspace applications
	access to internal databases
	access to external sources
Broadband data (data transmission over 384 kbps)	rapid file transfer
	high resolution photos
	remote operations
	mapping with Geographic Information System (GIS) layers
	mobile workspace applications
	access to internal databases
	access to external sources
Video (data transmission with tighter latency and coding requirements)	video transmission
	video streaming
	video call
Transversal services (extension of voice and data capabilities and performance)	extension of coverage
	extension of availability
	encryption tools
Challenging services (services enabled by the next generation of technologies)	proximity services
	augmented reality

III. ARCHITECTURE OF FUTURE PPDR SYSTEMS

A. Business Models

Migration to broadband systems can be examined using different business models for which CAPital EXpenditures (CAPEX) and OPerational EXpenditures (OPEX) can be split in different proportions. Three approaches were identified in [7] and described in [8]:

1) dedicated networks — a PPDR organization builds its own network infrastructure, or the build is done by a commercial operator based on a turnkey contract. The new network can be operated by the PPDR organization or by a commercial operator;
2) commercial networks — commercial operator(s) use public network(s) operated in order to provide PPDR services with a required Quality of Service (QoS);
3) hybrid networks — any mix of the above.

It means that the business model finally adopted by a PPDR body will significantly affect the whole process of acquiring a new network. It includes technical, financial-economic, and organizational issues. The hybrid network seems the most

Table II
LIST OF APPLICATIONS NEEDED BY PPDR

ID	Applications
APP1	Automatic (Vehicle) Location System (A(V)LS) data to Command and Control Centre (CCC)
APP2	A(V)LS data return
APP3	Video from/to CCC for following and intervention
APP4	Low quality additional feeds
APP5	Video for fixed observation
APP6	High quality additional feeds
APP7	Video on location (disaster or event area) to and from control room - high quality
APP8	Video on location (disaster or event area) to and from control room - low quality
APP9	Video on location (disaster or event area) for local use
APP10	Video conferencing operations
APP11	Non real time recorded video transmission
APP12	Photo broadcast
APP13	Photo to selected group (e.g., based on location)
APP14	Personal Information Management (PIM) synchronization in Personal Digital Assistant (PDA)
APP15	Mobile workspace (including public Internet)
APP16	Incident information download (text and images) from CCC to field units and netcentric working
APP17	Automatic Number Plate Recognition (ANPR) update hit list
APP18	Download maps with included information to field units
APP19	Command & control information including task management and briefings
APP20	Incident information upload (text and images) to CCC and netcentric working
APP21	Status information and location
APP22	ANPR or speed control automatic upload to database including pictures (temporally 'fixed' cameras and from vehicles)
APP23	Forward scanned documents
APP24	Reporting including pictures, etc.
APP25	Upload maps and schemes with included information
APP26	Patient monitoring - snapshot to hospital
APP27	Patient monitoring - real time monitoring to hospital
APP28	Monitoring status of security worker (drop detection, stress level, carbon monoxide, etc.)
APP29	Operational database search (own and external)
APP30	Remote medical database services
APP31	ANPR checking number plate live on demand
APP32	Biometric (e.g., fingerprint) check
APP33	Cargo data
APP34	Crash Recovery System (asking information on the spot)
APP35	Crash Recovery System (update to vehicles from database)
APP36	Software update online
APP37	GIS maps updates
APP38	Automatic telemetrics including remote controlled devices and information from static sensors
APP39	Hotspot on disaster or event area (e.g., in mobile communication centre)
APP40	Front-office and back-office applications, form filling online with back-office system, etc.
APP41	Alarming / paging
APP42	Traffic management system: information on road situations to units
APP43	Connectivity of abroad assigned force to local CCC
APP44	Unmanned Aircraft System (UAS) and Unmanned Ground Vehicle (UGV) control applications
APP45	Sensors on site

promising approaches but also the most challenging because two (or even more) networks have to be integrated in order to provide services in a seamless and efficient way. Typically it involves the setting up as a PPDR Mobile Virtual Network Operator (MVNO) over a commercial network(s) [9]. It was also noticed in [10] that hybrid solutions involve both dedicated specialized and commercial networks. This hybrid approach combines existing PPDR networks (e.g., TETRA) with a phased move to a common LTE infrastructure. A hybrid solution is also discussed in [11] as the most economic strategy where the dedicated network is operated in areas with high density of population, and services in rural areas are provided by commercial network(s).

B. System Architecture of Shared Radio Access Network

In this approach a broadband Radio Access Network (RAN) is managed by a Commercial Mobile Network Operator (CMNO) who shares it with the PPDR agency. A PPDR organization owns its 3G/LTE core network. Such a network is used for voice and data. However, mission-critical voice and narrowband data services remain on the Professional Mobile Radio (PMR) network already used, e.g., TETRA. The own core network enables the PPDR operator to have full control of the PPDR users with respect to their subscriptions, service profiles and service portfolio.

The network architecture is shown in Fig. 1. There are three sub-networks:

- 3G operated by CMNO,
- 4G operated by CMNO,
- TETRA/TETRAPOL operated by PPDR agency.

There are also a few components operated by the MVNO that are needed to provide access to the network resources. MVNO is the PPDR agency or an operator delegated by PPDR agency. In this strategy there are many commitments that have to be met by the MVNO and CMNO.

CMNO's 3G network consists of:

- RAN that composes of:
  - Node-B (NB) base stations,
  - Radio Network Controllers (RNC),
- core network where the most important components are:
  - Serving GPRS Support Node (SGSN),
  - Gateway Mobile Switching Center (GMSC).

MVNO's 3G core network is composed of:

- for Circuit Switching (CS):
  - Mobile Switching Center (MSC),
  - Visited Location Register (VLR),
  - Equipment Identity Register (EIR),
  - GMSC,
- for Packet Switching (PS):
  - Gateway GPRS Support Node (GGSN).

In 3G CS domain MVNO operates:

- CS GateWay (CS GW).

Common components for 3G CS and PS domains in the MVNO network are the following:

- Home Location Register (HLR),
- Short Messaging Service Center (SMS-C),
- Multimedia Message Service Center (MMS-C),
- Value-Added Service (VAS) platform.

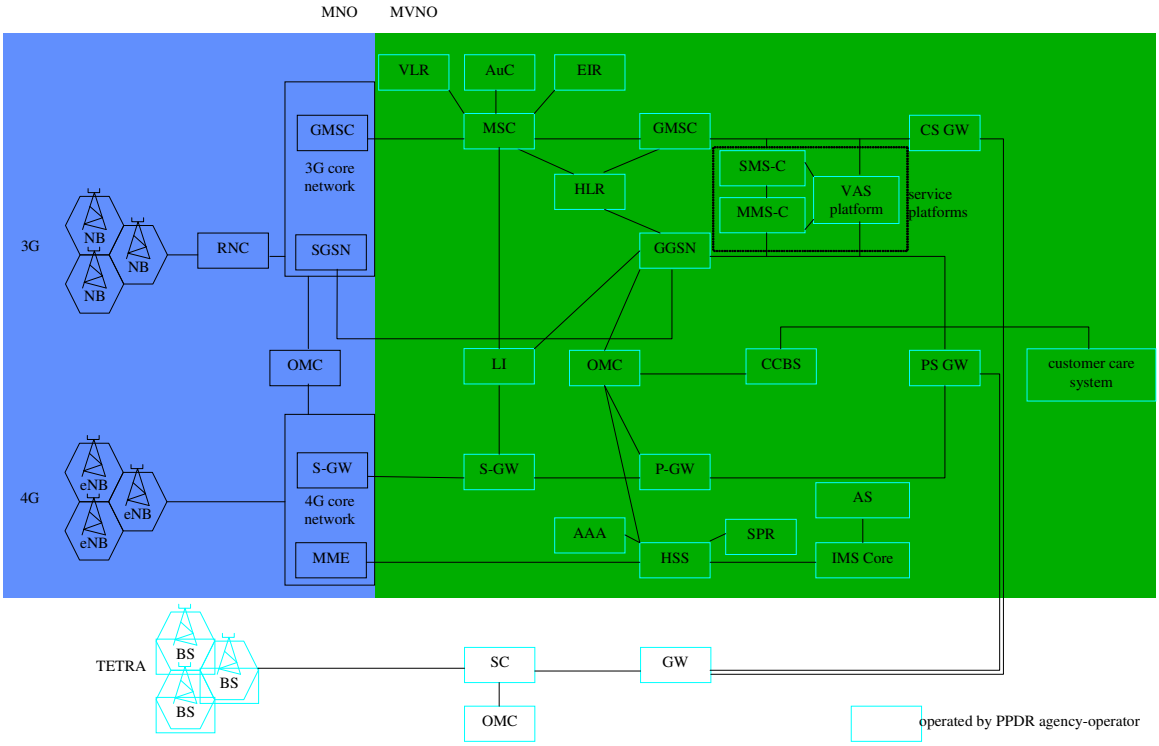


Figure 1. System architecture when PPDR operator is a Pure MVNO

CMNO's 4G network consists of:

- RAN that composes of:
  - evolved NB (eNB),
- core network where the most important components are:
  - Mobile Management Entity (MME),
  - Serving GateWay (S-GW).

MVNO's 4G core network is composed of:

- S-GW,
- Packet Date Network (PDN) GateWay (P-GW),
- Home Subscriber Server (HSS),
- Authentication, Authorization and Accounting (AAA),
- Subscription Profile Repository (SPR),
- IMS (Internet Protocol Multimedia Subsystem) Core,
- Application Servers (AS).

Moreover, there are some common elements for 3G and 4G sub-networks operated by MNO:

- backhaul network,
- backbone network,
- Operation and Maintenance Center (OMC).

At the MVNO site there are also some common elements for 3G and 4G sub-networks:

- Lawful Interception (LI),
- OMC,
- Customer Care and Billing System (CCBS),
- Packet Switch GateWay (PS GW) used by 3G PS domain and 4G.

MVNO who is a PPDR agency-operator governs:

- 3G CS and PS core,
- service platforms,
- 3G CS and 3G/4G PS gateways,
- 4G core,
- provision of services using IMS Core and AS,
- customer care system that allows managing PPDR entities' customers.

In TETRA/TETRAPOL sub-network there are:

- Base Stations (BS),
- Switching Center,
- interoperability GateWay (GW),
- OMC.

Distinction of network elements and processes handled by operators is given in Fig. 2. PPDR Pure MVNO is engaged in:

- O&M of core network,
- operation of Point of Interconnect (PoI),
- operation of LI,
- operation of different types of register (HLR/VLR, EIR, SPR),
- authentication, authorization and ciphering,
- operation of VAS platform including SMS-C, MMS-C, Internet access, IMS Core and ASs as well as Application Programming Interfaces (API) to service and content providers.

The remaining tasks are done by MNO. It includes operation of:

- RAN, backhaul and backbone,

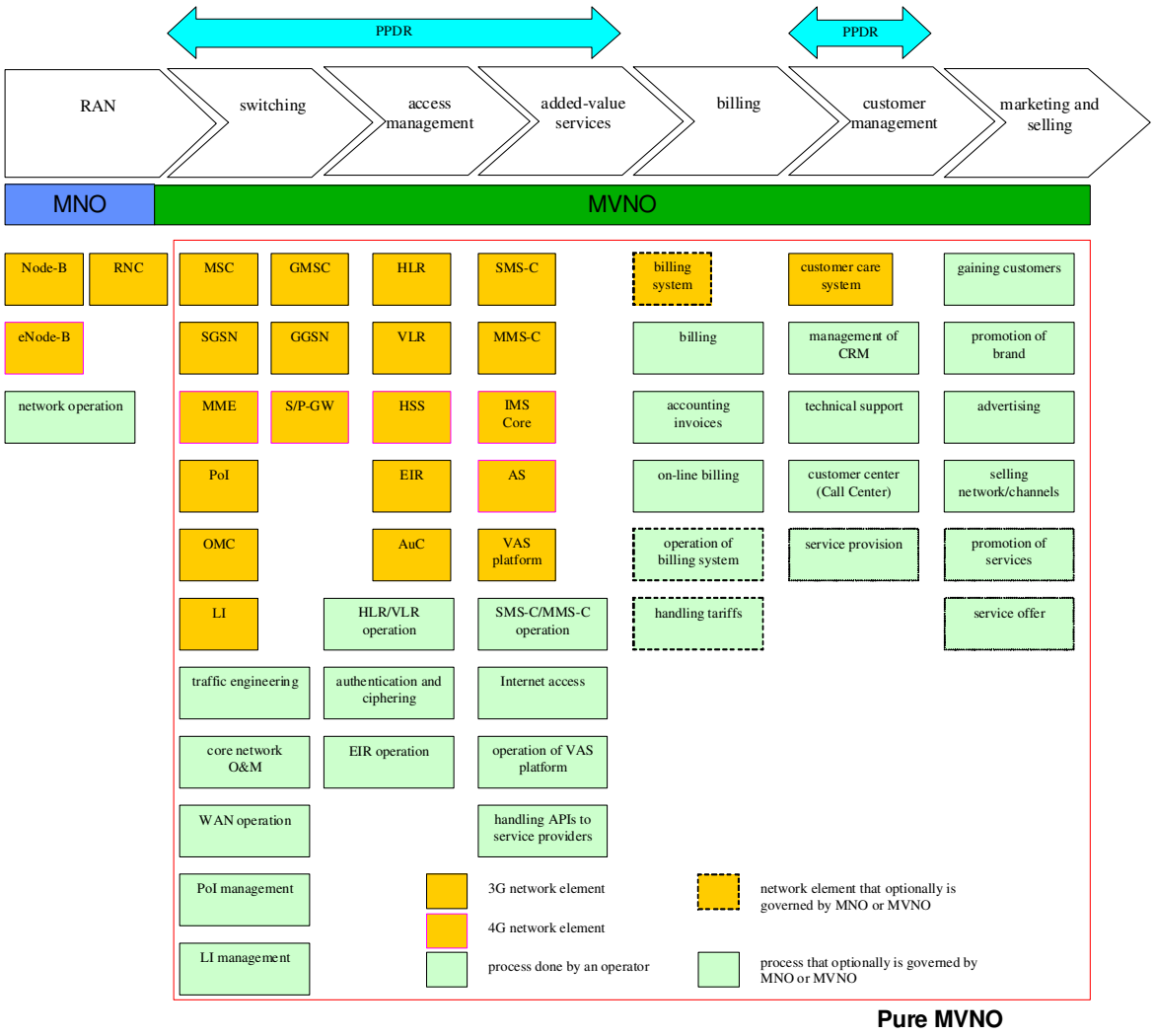


Figure 2. Value-added chain for Pure MVNO

- its core network.

Such a system architecture needs reliable interconnection and high QoS. The challenge in such an infrastructure is the provision of QoS based on Service Level Agreement (SLA). For the PPDR organization the quality requirements may be more stringent than those required for a public network. The cost implications to upgrade the mobile network may not be economically viable for a MVNO agreement with the PPDR organization.

However, in this approach there is the potential for PPDR agencies to obtain a level of independence from CMNO. A pure MVNO can ensure higher level of security and can deploy their own services quicker.

Potential sub-options to the presented architecture are hybrid models where the PPDR operator is:

- MVNO with shared RAN network and MNO with dedicated RAN network; as MNO the PPDR operator has own frequency carriers within relevant spectrum resources available nationwide; the own RAN infrastructure is built by PPDR agencies in some parts of a country to cover:
  - the most important/risky areas,
  - areas where there is no coverage from CMNOs used as hosts for virtualization;
- MVNO with shared RAN network but this sharing concerns all RAN components except frequency band; PPDR organisations have own frequency carriers within relevant spectrum resources available nationwide. These frequency channels are allocated to host's eNBs.

In all cases the MVNO agreement can be reached with one or more MNO host operators to increase coverage, availability and capacity. Advantages and disadvantages of this approach are summarized in Table III. In the above two cases a PPDR operator is partly MVNO so such a model affects both CAPEX and OPEX.

Regarding spectrum issues, this approach assumes that the 3G/LTE network operate in the bands licensed to the CMNO so the bands depend on the operator and the country. The solution also remains feasible for future spectrum bands dedi-

Table III
ADVANTAGES AND DISADVANTAGES OF SHARING RAN BY PURE MVNO TO PROVIDE BROADBAND SERVICES TO PPDR AGENCIES

Advantages	Disadvantages
Moderate CAPEX	Some boundaries on the transfer volume can be established by CMNO
Quicker deployment than building the network by own	Moderate OPEX
Mission-critical voice services are secured using PMR network	Low network availability during crisis events due to congestion in CMNO's infrastructure
In non critical cases RAN can be used for voice and data to increase overall capacity	Coverage depends on CMNO
Security can be hardened using own core network	No priority of services for PPDR agencies
Easier to deploy new services	Lack of resilience in RAN leading to low availability

cated for PPDR (e.g. proposed 700 MHz bands for harmonized Public Safety use). Moreover commercial operators may be interested in getting access to new PPDR spectrum that would be designated for PPDR use as the highest priority use but it could be available for commercial operation when not used by PPDR users. Such an approach needs, however, further investigation and strict control including immediate preemption of commercial users when PPDR spectrum is needed for designated users.

In the planning phase of this solution one has to take into consideration the potential for outages of the broadband network due to congestion during crisis events. It affects voice, message and data which are vital.

IV. UPGRADE OF PPDR NETWORKS TOWARDS BROADBAND SERVICES

Due to many factors the migration path to broadband PPDR networks will consist of a few steps. Nationwide TETRA/TETRAPOL networks were expensive but expenditures for broadband ones will be even greater because more base stations are required to cover the same area if higher frequency bands are used. On the other hand, some infrastructure elements (e.g. masts, server rooms, etc.) are available and can be re-used. Many other infrastructure elements will need to be upgraded (e.g. backhaul network) or even replaced (e.g. base stations). Moreover, this evolution has to be harmonized with other considerations such as LTE standardization to ensure equipment / networks support the requirements for Mission Critical Communications and with the release of frequency bands to be re-allocated for PPDR communication.

It seems that this evolution of PPDR networks will be based on several intermediate business approaches. The other assumption is that TETRA networks remain in use due to their maturity and great resiliency. However, the other option is to turn off the TETRA networks as soon as the broadband networks are operational to reduce overall costs (e.g., energy consumption and updates) borne by PPDR organizations. As shown in Fig. 3, 3G/4G broadband services provided by commercial MNOs can be used by PPDR agencies now. The evolution path can begin when 3GPP Release 13 is standardized and equipment is available to support Group Communication service and Mission Critical Push-To-Talk (PTT) one. Equipment might become commercially available after 2018It could happen after 2018.

The first step is for the PPDR organization to set up as a MVNO on a number of 4G networks to improve availability and reliability of services. This can be time consuming because many contracts have to be negotiated and signed as in the case of roaming.

The second step can begin when the 700 MHz band is released for PPDR mobile purposes. How this band will be arranged is still under international discussion by EU [12]-[13], International Telecommunication Union (ITU) and European Conference of Postal and Telecommunications Administrations (CEPT) [14]-[16], as well as by national regulation authorities [17]. In [12] a proposal was made to defer co-primary allocation of spectrum below the 700 MHz band to the mobile service until 2030 in order to give political and business reassurance for terrestrial television broadcasting and Program Making and Special Events (PMSE) applications. After 2025 a discussion can be reopened how this band can be allocated to inform stakeholders in advance before the deadline for safeguards of 2030. Now it seems more clear that further protection of the Ultra High Frequency (UHF) band for the production and ubiquitous delivery of audiovisual content is not needed after that deadline and mobile broadband services will be able to use it. At the national level the release of radio channel allocations is another issue. The expiry dates of existing broadcasting licenses in the UHF band are 2023 and 2026 for 48.5% and 72.8% of countries, respectively. That is why the second step of the migration has to be delayed. In this step the PPDR organization becomes an MNO with its base stations and frequencies deployed in hot-spots in limited geographic areas but it still maintains the *status quo* as a MVNO because the coverage of its own RAN is limited.

Finally, the PPDR MNO can operate dedicated radio channels in both commercial and own RAN networks. In this third step a close collaboration with at least one CMNO is still needed because it supports the access to its RAN.

In all steps the PPDR body can additionally reach national roaming contract(s) with CMNO(s) in order to ameliorate QoS, network coverage, and service availability. Such a CMNO is not the host for the MVNO but collaboration rules remain very similar.

In Fig. 3 mobility vs. capacity is shown for different technologies including TETRA, Universal Mobile Telephony System (UMTS), High Speed Packet Access (HSPA) and LTE/LTE-Advanced. The demand for higher speed services will limit the number of users that can be supported. The instant throughput also depends on the distance of the mobile from the base station and is controlled with adaptation of transmission parameters to fit to propagation conditions, e.g., Adaptive Modulation and Coding (ACM) mechanisms.

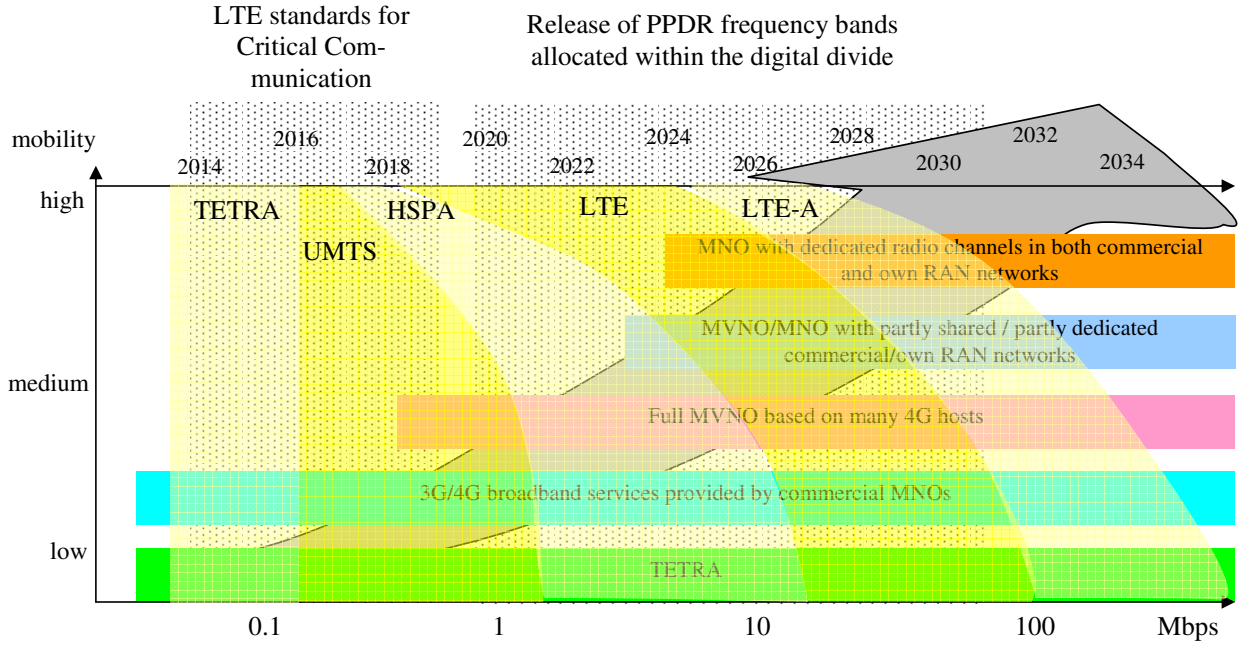


Figure 3. Migration path to broadband PPDR networks (LTE-A – LTE-Advanced).

V. Evaluation of Expenditures

The expenditures of such a migration process presented in Sec. IV were evaluated using a software tool [8] developed to compare different business models. The tool supports top-down and bottom-up analysis and takes into account a number of considerations including technical, financial-economic and organizational aspects.

In this example use case we have estimated the migration costs of a nationwide network for the energy sector in Poland. Needs defined by the energy sector that are related to the bearer services in the communication scenarios (cf. Sec. II) are presented in Table IV. In the case of major incidents the voice service is the key requirement as it provides instantaneous and the fastest way for personnel to exchange information. In the energy sector's network there are 111 000 terminals in operation and 50% of them will be replaced with LTE terminals during the migration phase up to 2025. Within step 2 and step 3 the energy sector will build 1600 eNBs. CAPEX and OPEX of each step along the migration path is shown in Fig. 4. The greatest CAPEX is in step 3 because the greatest number of base stations is built in this period. The energy sector has to pay license costs for reservation of radio channels in the 700 MHz band as well. The greatest OPEX is observed in step 1 due to expenditures related to operation and maintenance of the whole infrastructure delivered by the CMNO host.

Table IV
Communication needs matrix for energy sector in major incidents

Communication scenario	Voice	NB data	BB data	Video	Repeater
A	●	◑	○	○	⊙
B	●	⊙	○	○	⊙
C	●	○	○	○	⊙
D	◑	○	○	○	○
E	⊙	⊙	○	○	○
F	○	○	○	○	⊙
G	⊙	⊙	○	○	⊙
H	○	⊙	○	○	⊙

Notation of needs: ●High, ◑Medium, ⊙Low, ○Not required

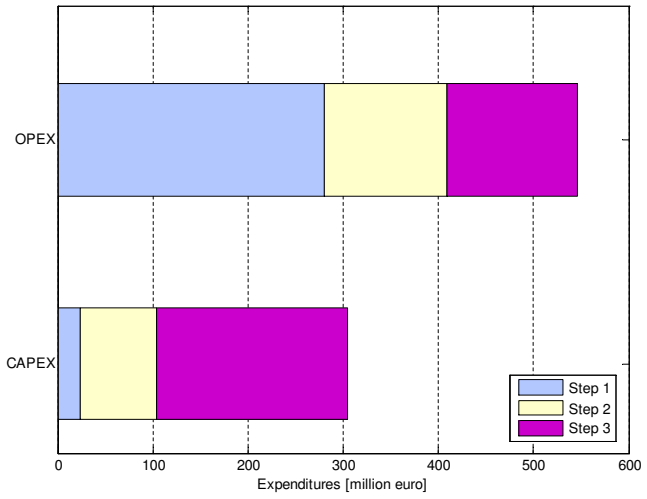


Figure 4. CAPEX and OPEX breakdown for all steps in the migration path

VI. Conclusion

In this paper we have presented several issues related to future migration of PPDR communication networks to broadband. PPDR organizations would like to benefit from the current evolution in mobile technologies to use broadband

services. One can envisage that future PPDR networks will be based on LTE (LTE-Advanced) technology but existing narrowband networks (e.g. TETRA) will still be required due to their maturity and proven provision of mission critical voice communications. For example, German BOSNet network based on TETRA standard is planned to be fully deployed this year. TETRA networks will be also built by the energy sector's companies in Poland. Nevertheless, LTE technology has to be significantly developed in order to meet the requirements of Public safety organizations and work is currently ongoing within the 3rd Generation Partnership Project (3GPP) as well as by TETRA and Critical Communications Association (TCCA). The main benefit from the deployment of LTE broadband networks is the provision of video. In the UK a dialogue with manufacturers began last year in Great Britain to inform the replacement process of the current TETRA-based Airwave network. The plan is for a commercially operated new 4G LTE-based mission-critical network for public-safety and other governmental organizations.

One of the crucial aspects of the migration path is selecting a business model that meets the requirements of a broadband network defined by the PPDR organizations. The provision of a new own network can involve considerable expenditure and hybrid approaches can significantly reduce costs. However models based on being a MVNO require detailed planning, negotiations and acceptable SLAs to both parties. Other options where the PPDR organizations are a MVNO and also roll-out a limited own network to ensure coverage and/or capacity in specific geograhic areas may be attractive if suitable spectrum can be made available.

There may be the potential to use the 700 MHz band but its use depends on existing reservations in each country. Also the re-allocation of the spectrum may take several years and will vary by country. Therefore, the process of evolution towards broadband networks will certainly take a few years.

ACKNOWLEDGMENT

H. G. gratefully acknowledges Jacek Jarzina from TELECOM and Sławomir Fryska from YAGI-FRYSKA for interesting and valuable discussions on deployments of PMR networks.

REFERENCES

[1] C. Lucente, "700 MHz spectrum requirements for Canadian public safety interoperable mobile broadband data communications," Martello Defence Security Consultants Inc., 2011.
[2] "Comments of the City of New York in the matter of additional comment sought on public safety, homeland security, and cyber security elements of the National Broadband Plan (NBP) Public Notice no. 8," 2009.
[3] "User requirements and spectrum needs for future European broadband PPDR systems (Wide Area Networks)," ECC Report 199, 2013.
[4] J. S. Marcus, J. Burns, V. Jervis, R. Wählen, K. R. Carter, I. Philbeck, and P. Vary, "PPDR Spectrum Harmonisation in Germany, Europe and Globally," WIK-Consult, Report prepared for BMWI Germany, 2010.
[5] "Updates Broadband Speed Benchmark to 25 Mbps to Reflect Consumer Demand, Advances in Technology," News, Federal Communications Commission, 2015.
[6] "User Requirement Specification; Mission Critical Broadband Communication Requirements," ETSI TR 102 022-1, 2012.
[7] CEPT ECC Radio Spectrum for Public Protection and Disaster Relief (PPDR) working group, FM49(14)008rev1 Network Types part (Draft ECC Report B).
[8] H. Gierszal; K. Pawlina, P. Tyczka, K. Romanowski D. Lavaux, and C. Katsigiannis, "Business models and multi-domain analysis for acquiring broadband PPDR systems," 10th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 8-10 October 2014, Larnaca, Cyprus, pp. 334-340, DOI 10.1109/WiMOB.2014.6962191.
[9] T. Bassayiannis, *Mobile Virtual Network Operator (MVNO)*, MBIT Thesis, Athens Information Technology, 2008.
[10] S. Forge, R. Horvitz, C. Blackman, "A study on the use of commercial mobile networks and equipment for "mission-critical" high-speed broadband communications in specific sectors," SMART 2013-0016, SCF Associates Ltd., Stakeholder Workshop, 30 April 2014.
[11] J. Vinkvist, T. Pesonen, and M. Peltola, "Finland's 5 steps to critical broadband," RadioResource International, 4/2014.
[12] P. Lamy, "Results of the work of the high level group on the future use of the UHF band (470-790 MHz)," Report to the European Commission, 2014.
[13] "Draft RSPG Opinion on a long-term strategy on the future use of the UHF band (470-790 MHz) in the European Union," EC RSPG 14-585(rev1), 2014.
[14] "Harmonized technical conditions for mobile/fixed communications networks (MFCN) in the band 694-790 MHz including a paired frequency arrangement (Frequency Division Duplex 2x30 MHz) and an optional unpaired frequency arrangement (Supplemental Downlink)," ECC Decision (15)01, 2015.
[15] "Report A from CEPT to the European Commission in response to the Mandate: "To develop harmonized technical conditions for the 694-790 MHz ('700 MHz') frequency band in the EU for the provision of wireless broadband and other uses in support of EU spectrum policy objectives". Provisional lower band edge subject to precise definition within the scope of this Mandate", 2014.
[16] "Long Term Vision for the UHF broadcasting band," ECC Report 224, 2014.
[17] M. J. Grzybkowski, "Systemy PPDR w przestrzeni widmowej II dywidendy cyfrowej" (PPDR systems in the view of II digital divide), Przegląd Telekomunikacyjny (Telecommunications Review), no. 4/2015, pp. 477-480, DOI: 10.15199/59.2015.4.92 [in Polish].

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 159–164

DOI: 10.15439/2015F37
ACSIS, Vol. 6

Reconfigurable FPGA-based embedded Web services as distributed computational nodes

Robert Brzoza-Woch *, Piotr Nawrocki †
*AGH University of Science and Technology,
al. A. Mickiewicza 30, 30-059 Krakow, Poland
e-mail:rabw@agh.edu.pl
†e-mail:piotr.nawrocki@agh.edu.pl

***Abstract*—In this article we propose a concept for an experimental class of control devices that use both a microcontroller unit (MCU) and a field-programmable gate array (FPGA) circuit. These devices can provide the functionality of full-featured Web services that are compliant with the Service-Oriented Architecture (SOA) paradigm. Despite the fact that FPGA circuits are more expensive than consumer-grade MCUs, they potentially offer much more computational power. In scenarios in which FPGA computational power is required on demand and for short periods only, a large part of such resources might, however, remain unused or disabled. Thus we propose a system architecture and software infrastructure that simplify the utilization of temporarily unused resources for performing various tasks that can be offered as Web services on a commercial basis.**

I. Introduction

FPGA-based hardware Web services have already been implemented and described (refer to [1] and [2]). Their embedded nature allows developers to easily adapt those services to actively interact with their environment, e.g. to acquire real-world measurement data or control various actuators. Such entities can be called *environment-aware* Web services in contrast to classical Web services that work on remote physical or virtual machines. Despite the fact that environment-aware Web services may be implemented using much less expensive MCUs and sequential code, programmable hardware may perform better where very intensive computational tasks are involved.

In our solution we propose that environment-aware Web services can be reconfigured in order to exploit the potential of their temporarily unused logic resources. At times of lower utilization they can be reconfigured to offer their spare resources as additional data-processing Web services. Whenever a more intensive processing task is to be performed, their resources can be employed back to provide the device's original functionality. This idea can also be applied to regular devices that offer no Web service compliance. In the latter case, however, we would lose some useful features such as interoperability or the ability to utilize the management software tools already available, etc.

There are common scenarios where considerable computational power is required only on demand for a short period. For example, in an industrial process temperature measurement results are collected for many hours to finally update a local numeric prediction module. During the time of collection, the acquisition device may be idle or in sleep mode or, if it uses FPGA, a large part of its logic may be disabled. By using the dynamic reconfiguration technique for FPGA, we can change the device's configuration depending on momentary needs, e.g. to adapt to changing environment conditions or the current context.

FPGA-based environmental-aware Web services may also be employed to perform control tasks in smart home automation systems. The less demanding the control task, the more resources can be assigned to perform "idle" computations. For example, an intelligent water heater may periodically compute predicted hot water usage and perform computations for a commercial Web service during its "spare" time.

In this paper we describe the concept for a system that uses FPGA-based Web services to perform such dual operation. In Section II we present the current state of the art in related fields. Then in Section III we describe the architecture of sample FPGA-based Web services and discuss the architecture of the entire system. In Section IV we introduce the service management and integration mechanism, which is responsible for providing the essential functionality of the system and ensuring its security. Finally, in Section V we conclude our work and discuss the planned evolution of our solution.

II. Related work

Web service implementations using FPGAs can be found, however such publications are relatively rare. In [3], the authors propose a Web server architecture that is implemented in FPGA. The evaluation of the system confirms that hardware-favored architecture brings higher throughput, lower power consumption and the full functionality of a stand-alone Web service. Such good results are achieved thanks to the execution of Web services directly on the FPGA without using an additional operating system. The authors conclude that by utilizing reconfigurable hardware (FPGA) in the area of cloud computing it is possible to improve performance and optimize operating costs.

Important research in this field is described in [4]. The authors present a reconfigurable architecture for Web service implementation. The important features of the system presented are:

- high overall performance because of the very low response time and potentially high processing power;
- a reconfiguration ability which allows the system to be updated to meet new requirements.

Both advantages mentioned result from the use of the FPGA technology. The platform supports the SOAP protocol and is able to auto-register into a UDDI server. Even more interestingly, the platform presented works without any embedded microcontroller (such as NIOS-II) yet it is partially implemented using the Handel-C language. The overall performance results are good because the platform has a lower response time (a minimum of 0.5 ms) than a PC running the same service (a minimum of 2 ms). A disadvantage of the solution described is its very simple functionality—in the configuration presented, it only provides a Wake on LAN service for computers within a local area network.

Another FPGA-based Web service implementation is described by C.E. Chang in [5]. It is a RESTful Web service designed to perform simple control tasks for home appliances. The service functionality is rather minimalistic and its implementation has less features than the system described in [4], e.g. it supports static IP assignment only (no DHCP), the TCP/IP buffer size is limited to just a single packet or 576 bytes, and only one TCP connection can be accepted at a time.

The possibility of using reconfigurable hardware and Web Service technology within the framework of the concept of the Internet of Things (IoT) is described in [6]. The authors of that paper discuss the prospects of reconfigurable hardware solutions in the area of enterprise applications. They present requirements for reconfigurable computing solutions and argue that this type of platform can assist the performance of business processes. They also estimate that reconfigurable computing platforms will play a key role in connecting two worlds: the “Internet of Services” and the “Internet of Things”.

In [7], the authors note the need for connecting the concept of IoT with service-oriented methodology. They suggest the use of RESTful Web services due to their popularity and lightweight nature. In order to utilize the RESTful concept in the IoT domain, the article proposes an architecture composed of six levels. The authors also demonstrate how to invoke RESTful Web services from the IoT and publish a Business Process Execution Language (BPEL) process as a RESTful Web service.

Important areas that utilize the IoT and Web service concepts are the Smart Home and Smart Building. In order to provide IoT services in Smart Home/Smart Building environments, in [8] the authors propose a Web-of-Objects platform in the IoT service environment. This platform has been designed in order to create user-centered IoT services. In addition, complex services can be developed by combining elements of existing Web services.

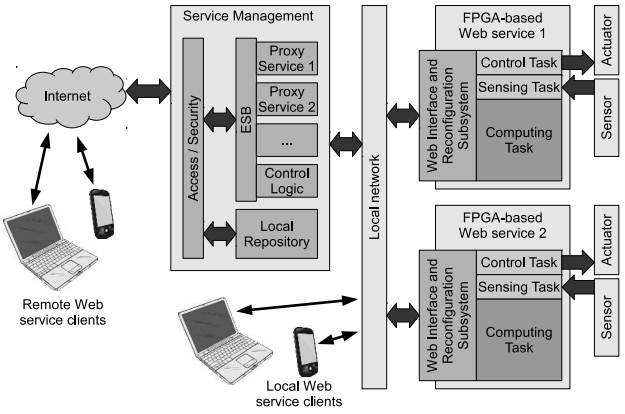


Fig. 1. General concept of networked reconfigurable FPGA-based Web services.

III. Hardware infrastructure and system architecture

As stated in Section I, SOA-compliant Web services may be implemented not only on classical server farms, but also on embedded devices, obviously including FPGA-based ones. During the course of our research we developed several FPGA-based Web services on various hardware platforms. One of the most versatile hardware platforms we developed was described in detail in our previous publications: [1] and [2]. Each platform was equipped with a high-end Stratix-II FPGA chip and multiple blocks of synchronous dynamic random access memory (SDRAM) and was able to support built-in as well as external sensors and actuators. In this article we describe how we experimented with the deployment of those platforms within dynamically managed and reconfigurable distributed systems.

A. System architecture

The architecture of the distributed control computing system is shown in Fig. 1. In that scenario, FPGA-based Web services perform several tasks. First, they provide part of the Web service interface and perform their default control and sensing tasks that require only a fractional amount of their logic resources while their spare resources are assigned to perform a *Computing Task*. We assume that the computing task is required for the process being supervised only during certain periods, for example after a sufficient amount of data has been collected – this is quite a common scenario in computer systems. During the service’s lower activity periods, its resources can be assigned to perform a completely independent task thanks to the massive parallel capabilities of the FPGA technology.

FPGA-based Web services can be conveniently interconnected within a local area network or may operate in a virtual private network (a VPN, which is logically equivalent to local area network operation). At this basic level, data transmission security can be ensured either using the Wi-Fi Protected

Access (WPA) technology for wireless entities or by using wired connections (it is fair to assume that an attack on a building's backbone network is equally invasive as tampering with physical devices). Depending on security requirements, we may or may not allow trusted clients to have unrestricted access to FPGA-based Web services in the local area network. Access restrictions can be introduced using various techniques, even very simple ones: e.g. by means of Media Access Control (MAC) address filtering.

More sophisticated features are available when using the *Service Management* subsystem shown in Fig. 1. The management subsystem runs on a stand-alone computer or on server infrastructure and it provides *Proxy Services* for each physical or logical Web service on the network. Proxy Services facilitate access to local embedded services by increasing the number of simultaneous client connections and introducing the Enterprise Service Bus (ESB). ESB significantly increases the solution's scalability and accessibility from mobile devices. Separate services are connected to the ESB to provide control logic and access to the local service repository. The *Control Logic* service on ESB is the core of our concept. It decides whether an FPGA-based Web service should be reconfigured and which of these services can be offered as spare computational resources to the outside world. The logic may also advertise available resources to service brokers. As computational services may not be always available on the programmable hardware (during idle periods only), the control logic should be able to move the execution of tasks from FPGA-based services to its internal software and back to the FPGA after hardware resources have been released again. At lower level, this functionality can be easily implemented on various existing FPGA platforms with additional reconfiguration hardware. Another challenge concerns the algorithms that support decision whether the functionality should be moved to or from FPGA to the control logic. Those algorithms may vary from a single busy-idle state detection to more advanced solutions which also consider potential time and energy costs of the task moving operation. That functionality, however, is still under development and is to be introduced in future versions of the system.

B. FPGA-based Web service architecture

In order to cooperate within the system presented, each FPGA-based embedded Web service should follow some particular architectural recommendations. The most trivial issue with the FPGAs' remote reconfiguration is that most common FPGAs completely lose their previous functionality during the process of reprogramming their internal memory. This is the reason why all FPGA-based Web services should be implemented as *hybrid* devices. A simple way to ensure uninterrupted operation in this case is to implement part of the communication interface and reconfiguration logic on a different chip. In the most trivial case, the reconfiguration logic can be implemented on a popular inexpensive microcontroller unit (MCU) with network communication capabilities as shown in Fig. 2.

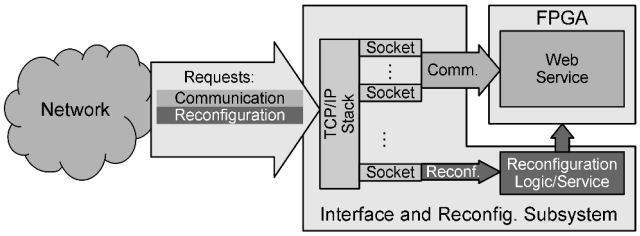


Fig. 2. The idea of multiplexing network hardware between the FPGA-based Web service and the reconfiguration subsystem.

In practice, a relatively simple and effective solution was to introduce multiplexing at the network socket level by implementing communication functionality up to the transport layer of the Open System Interconnection (OSI) model. Depending on the available resources of the reconfiguration subsystem, performance constraints may occur. In practical implementations it was not the case, for two reasons. First, FPGA-based Web services tend to process data locally whenever possible and transmit processing results only. For example, a smart camera, which we have previously implemented (refer to [2]), runs a motion detection or object classification service. Instead of transmitting an entire video frame, the camera sends mainly coordinates for which it detected motion or the identification number of the object recognized. Secondly, each hardware service is represented by its "mirror" proxy service in the Management subsystem. This allows the hardware service to be exposed by a machine that has better networking capabilities than a typical embedded system. Reconfiguration data also do not have to be transmitted for each functionality, but can be cached locally on each node instead as described further in this section.

The reconfiguration logic of FPGA-based Web services can be exposed either as a constant method available for each service or else can use an application-specific protocol. Both these options have their advantages. Exposing reconfiguration capability as a service makes for a very clean system design and ease of integration as a proxy in the Management block. However, using a simple custom protocol may offer better performance because it does not involve the high-level protocol overhead that is unavoidable when using e.g. SOAP. A simple but efficient enhancement of the reconfiguration subsystem is the introduction of mass storage capability, e.g. in a form of a popular secure digital (SD) card or another form of non-volatile memory. This allows the system designer to cache the most common configuration data locally for each hardware Web service. In our solutions we used SD and SD High Capacity (SDHC) cards as well as DataFlash memory manufactured by Atmel.

Network communication can be implemented in several ways. The obvious choice for wired nodes is to use an Ethernet connection for the easiest possible interoperability with the network infrastructure already available. In our solutions, we

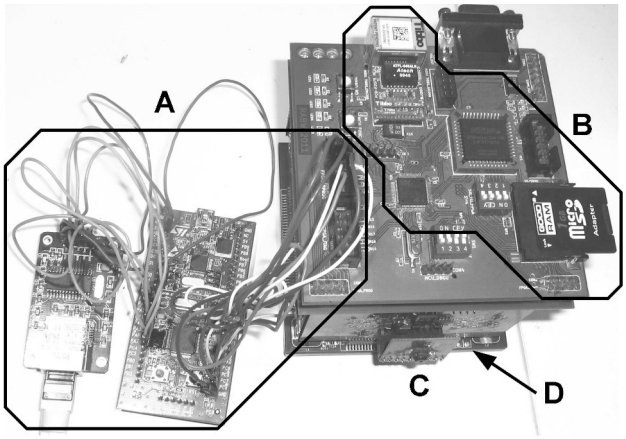


Fig. 3. View of the hardware part of sample implementations: the newly developed inexpensive remote reconfiguration and communication subsystem (RCS) based on the ENC28J60 and a mid-end STM32 MCU (A), the previously developed RCS based on the EM1206 network module, an STM32 MCU and CPLD, the embedded image sensor (C), and the Stratix-II FPGA module which provides Web service functionality (D).

initially tested in practice the operation of EM1206 network modules by Tibbo. These offer very simple programming capabilities thanks to their high autonomy and can be connected to any device (either an FPGA or a microcontroller) that supports asynchronous serial communication. Another simple solution is to use the ENC28J60 chip, which provides physical layer (PHY) capabilities and is equipped with a synchronous serial interface. Upper layers should then be implemented in microcontroller software. More demanding applications would require the utilization of a PHY chip with a fast Media Independent Interface (MII). Wireless network access can be provided using a Wi-Fi chip or module, e.g. ESP8266. It is a very inexpensive and easy-to-use module which became very popular recently. The ESP8266 makes it possible to connect virtually any embedded device to a Wi-Fi network and, importantly, has basic connection security measures already implemented. In our research we developed multiple FPGA-based Web services using EM1206 (wired) and ESP8266 (wireless) modules. Now we continue to further develop our solutions using other alternatives, mainly revolving round medium-power microcontroller (such as Cortex-M3) with external PHY chip. Fig. 3 shows a sample inexpensive hardware with ENC28J60 PHY and STM32F100 board that we have used in our development.

IV. Service Management Subsystem Details

The considerable computing capabilities of FPGA-based Web services make it worthwhile to enable them to interact and integrate with other systems. This type of solution, integrating independent but compatible services, lies at the foundation of the SOA concept and the Web service technology is one of the components of this concept. SOA may take advantage of service orchestration, which defines the model of cooperation between services. Within the framework of orchestration, there is a single parent process that manages the interoperability between services. The implementation of such a process, which allows for the provision of management and supervision services, is achieved via the Enterprise Service Bus (ESB), which specifies an intermediate layer that enables the integration of services. In order to enhance the processes of composing and configuring FPGA-based Web services, Proxy Service is created depending on current needs. This mechanism allows for the provision of FPGA-based Web service functionality using the ESB. The main advantages of this solution are standardization, scalability, reliability and manageability. The Proxy Service, which is implemented in Java, exposes the functionality of an FPGA-based Web service in an ESB container and allows it to better integrate with other services. Most Java implementations of ESB use the OSGi platform that allows, inter alia, the re-use of components, easy deployment and the ability to dynamically update components.

TABLE I
RESTful and WS* service comparison.

RESTful	WS-* (SOAP)
Lightweight architectural style	"Heavy-weight" XML standard
Description of the service in WSDL 2.0	Description of the service in WSDL
Format agnostic (XML, JSON, HTML, etc.)	Requests and responses are well structured (SOAP)
Problems with reliable messaging and security	Security mechanism possible (WS-Security)

In order to implement FPGA-based Web services and ensure their cooperation with the Proxy Service mechanism we could use two approaches: WS-* using SOAP and Representational State Transfer (RESTful) Web services built on the basis of HTTP. These are briefly compared in Table I. The first method describes the functionality of the service using the Web Service Description Language (WSDL), and communication with the service is implemented using the SOAP and HTTP protocols. The WS-* concept was originally developed for interoperability between enterprise applications. A lighter version of the WS-* was developed, named Devices Profile for Web Services (DPWS), in which services are directly associated with the equipment. The main DWPS area of application is home and industrial automation systems ([9]). In the second approach, RESTful services are identified by a Uniform Resource Identifier (URI) and the HTTP protocol is used to access the resources thus defined. The RESTful approach can be used to install services on smart devices. Some studies such as [10] suggest that for smart devices used in the development of IoT applications, a more suitable method for the provision of services is RESTful. On the other hand, the WS-* concept is more appropriate for applications requiring an adequate level of security and QoS. Therefore the authors of this article decided to choose the WS-* approach due to the potential application of the solution developed in systems that require reliable and secure data delivery.

After further analysis, we decided to use WS * and OSGi containers as our ESB implementations. Integration and ser-

vice management have been achieved through the exposition of FPGA-based hardware Web services using the Proxy Service—a service engine that is implemented in Java. The Proxy Service has been deployed within the ESB (OSGi container). The functionality of the service can also be exposed directly in a local area network as a Web service reachable using SOAP communication standards. In both cases, the interface can be specified in WSDL.

In our solution, each of the FPGA-based hardware Web services is described by an appropriate WSDL file. For each such service, it is possible to generate one corresponding Proxy Service with methods identical to those of the original service. All operations on a Proxy Service are delegated through standard WS requests (SOAP) to the server appropriate for the target device. Recently we have also developed experimental REST-based implementations of the embedded Web services because of the REST's simplicity and smaller communication overhead.

The automated generation process of a Proxy Service requires several steps:

1) generating a Proxy Service as an ESB adapter;
2) looking up a matching service reference in repositories;
3) downloading a WSDL file that describes the service's interface;
4) creating a hardware service interface in the Proxy.

Afterwards, the FPGA-based hardware Web service is available and ready to be used in enterprise class information systems.

The functionality of the FPGA-based hardware Web service is accessible through the Proxy Service mechanism thanks to mediation between the SOAP binding component and the OSGi Remote service binding component such as the R-OSGi (see [11]) or ECF Remote Services. The use of such ESB implementations as Fuse ESB 4.3, Apache ServiceMix 4.3.0 or Apache Felix (installed on mobile devices supporting the Android system) enables direct access to these services through the OSGi environment.

In addition to providing the functionality of the FPGA-based hardware Web service, each Proxy Service provides an interface that also allows for the remote reconfiguration of the FPGA device and provides completely new functionality. Details of the reconfiguration process are presented in Section III.

A very important issue in the integration and management of FPGA-based hardware Web services is the discovery process. All FPGA-based hardware Web services use network communication modules that provide a multi-socket TCP/IP stack. Some sockets have fixed roles and the rest are used for clients' connections. One socket is reserved for discovery purposes, enabling the FPGA-based hardware Web service currently running to be found within the local area network. There is a local (service) repository responsible for maintaining a directory of FPGA-based hardware Web services, detecting new services connected to the system and managing service configuration. The prototype service repository used in our solution utilizes the LDAP (Lightweight Directory Access Protocol) for FPGA-based hardware Web service inventory management purposes. Each device in the system is obliged to send beacon signals periodically. These beacon signals are sent as UDP datagrams to a multicast destination IP address and received by the service repository. Contents of each beacon message are generated in the internal logic of each individual FPGA-based hardware Web service and sent through the network communication module, which then transmits them using UDP.

An internal user who intends to interact with an FPGA-based hardware Web service browses the service repository in search of the service's details such as the invocation point and the signature of the operations supported. Upon finding them, the client application can connect to the hardware service directly. An external user can interact with an FPGA-based hardware Web service (find and run it) through the Internet and ESB.

A significant aspect of using FPGA-based Web services concerns the construction of a proactive system. Both protocols investigated (RESTful and WS*) have request/response characteristics that do not support the generation of asynchronous notifications to clients. The fact that the classic approach to Web Service technology has been selected implies that we cannot implement such a notification mechanism in our solutions. Notwithstanding this, in the concept developed we have the ability to send notifications on service status via the local (service) repository that is found under a well-known address, which acts as a proxy that performs notification operations. In the solution proposed, FPGA-based Web services automatically send registration information to the repository. The same method can be used to send information about service status to the repository. In this case, a separate notification mechanism can be implemented in the repository. Such a solution can be perceived as a step towards event-driven services.

During the design and implementation of our chosen solution, we also considered issues related to security. It was assumed that the local area network, which is not connected directly to the Internet, is safe. If secure communications in the local area network are required, SOAP and REST can be used over HTTPS. If wireless technology is used, security is provided by standard mechanisms such as Wi-Fi Protected Access (WPA / WPA2) as implmented in the ESP8266 modules that we use, Wireless Intrusion Prevention Systems (WIPS) or Wireless Intrusion Detection Systems (WIDS). User access from the Internet to FPGA-based hardware Web services is possible only through the ESB, where various user authorization and authentication mechanisms can be implemented through secure communication channels such as the HTTPS protocol.

V. Conclusions and future work

In this article, we present a concept for a better utilization of spare FPGA resources by employing them to perform independent computational tasks. We apply this approach to FPGA-based embedded and environment-aware Web services compliant with the SOA paradigm. Additional functional

modules have to be provided for each service and particular architectural guidelines have to be followed, which we present in this paper as a reference. We attempt to keep additional hardware costs as low as possible. Initially, we applied the concept presented to the previously developed FPGA hardware-software platform designed to run various Web services. Future development goals include 1) automatic service advertising (which is related to the issue of service repository [12]); and 2) developing or adapting available algorithms which would allow us to automatically move computations between FPGA-based Web services and the service management subsystem to ensure uninterrupted Web service operation.

ACKNOWLEDGMENT

The research presented in this paper was partially supported by the European Union in the scope of the European Regional Development Fund program no. POIG.01.03.01-00-008/08, the National Centre for Research and Development (NCBiR) under Grant No. PBS1/B9/18/2013 and by the Polish Ministry of Science and Higher Education under AGH University of Science and Technology Grant 11.11.230.124 (statutory project).

REFERENCES

[1] A. Ruta, R. Brzoza-Woch, and K. Zieliński, "On fast development of FPGA-based SOA services—machine vision case study," *Design Automation for Embedded Systems*, vol. 16, no. 1, pp. 45–69, 2012. doi: 10.1007/s10617-012-9084-z. [Online]. Available: http://dx.doi.org/10.1007/s10617-012-9084-z

[2] R. Brzoza-Woch, A. Ruta, and K. Zieliński, "Remotely reconfigurable hardware-software platform with web service interface for automated video surveillance," *Journal of Systems Architecture*, vol. 59, no. 7, pp. 376 – 388, 2013. doi: 10.1016/j.sysarc.2013.05.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138376211300074X

[3] J. Yu, Y. Zhu, L. Xia, M. Qiu, Y. Fu, and G. Rong, "Grounding high efficiency cloud computing architecture: HW-SW co-design and implementation of a stand-alone web server on FPGA," in *Applications of Digital Information and Web Technologies (ICADIWT), 2011 Fourth International Conference on the*, Aug 2011. doi: 10.1109/ICADIWT.2011.6041412 pp. 124–129.

[4] S. Cuenca-Asensi, H. Ramos-Morillo, H. Lloren-Martinez, and F. Macia-Perez, "Reconfigurable architecture for embedding web services," in *Programmable Logic, 2008 4th Southern Conference on*, March 2008. doi: 10.1109/SPL.2008.4547742 pp. 119–124.

[5] C. Chang, F. Mohd-Yasin, and A. Mustapha, "An implementation of embedded RESTful Web services," in *Innovative Technologies in Intelligent Systems and Industrial Applications, 2009. CITISIA 2009*, July 2009. doi: 10.1109/CITISIA.2009.5224244 pp. 45–50.

[6] M. Middendorf and B. Scheuermann, "Perspectives of extending runtime reconfigurable computing to the enterprise application domain," in *Industrial Informatics (INDIN), 2010 8th IEEE International Conference on*, July 2010. doi: 10.1109/INDIN.2010.5549416 pp. 266–273.

[7] L. Zhang, S. Yu, X. Ding, and X. Wang, "Research on IOT RESTful web service asynchronous composition based on BPEL," in *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on*, vol. 1, Aug 2014. doi: 10.1109/IHMSC.2014.23 pp. 62–65.

[8] Y. Kim, S. Lee, Y. Jeon, I. Chong, and S. H. Lee, "Orchestration in distributed web-of-objects for creation of user-centered iot service capability," in *Ubiquitous and Future Networks (ICUFN), 2013 Fifth International Conference on*, July 2013. doi: 10.1109/ICUFN.2013.6614920. ISSN 2165-8528 pp. 750–755.

[9] F. Jammes and H. Smit, "Service-oriented paradigms in industrial automation," *Industrial Informatics, IEEE Transactions on*, vol. 1, no. 1, pp. 62–70, Feb 2005. doi: 10.1109/TII.2005.844419

[10] D. Guinard, I. Ion, and S. Mayer, "In search of an internet of things service architecture: REST or WS-*? a developers' perspective," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, A. Puiatti and T. Gu, Eds. Springer Berlin Heidelberg, 2012, vol. 104, pp. 326–337. ISBN 978-3-642-30972-4. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-30973-1_32

[11] J. S. Rellermeyer, G. Alonso, and T. Roscoe, "R-OSGi: Distributed applications through software modularization," in *Proceedings of the ACM/IFIP/USENIX 2007 International Conference on Middleware*, ser. Middleware '07. New York, NY, USA: Springer-Verlag New York, Inc., 2007. doi: 10.1007/978-3-540-76778-7_1 pp. 1–20. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-76778-7_1

[12] P. Nawrocki and A. Mamla, "Distributed web service repository," *Computer Science*, vol. 16, no. 1, pp. 55–73, 2015. doi: 10.7494/csci.2015.16.1.55. [Online]. Available: http://journals.agh.edu.pl/csci/article/view/1205

4th International Conference on Wireless Sensor Networks

A FEW years ago, the applications of WSN were rather an interesting example than a powerful technology. Nowadays, this technology attracts still more and more scientific audience. Theoretical works from the past, where WSN principles were investigated, grew into attention-grabbing applications practically integrated by this time in a real life. It could be said, that countless application fields, from military to healthcare, are already covered by WSN. Together with this technology expansion, still new and new tasks and interesting problems are arising. Simultaneously, such application actions stimulate the progress of WSN theory that at the same time unlocks new application possibilities. The typical examples are developments within the "Internet-of-Things" field as well as advancements in eHealth domain with WBAN IEEE 802.15.6 standard progress.

Wireless sensor networks, as the spatially distributed networks consisted of a number of relatively simple, low-cost, low-power components interconnected mutually, provide quite wide application portfolio for different branches of economy. As the main examples could be mentioned military, industry, transport, agriculture and healthcare. However, in the near future, even stronger expansion of WSN application assortment is expected. In order to make this expansion possible, it is necessary to continually work on the solving of typical questions/problems related to the WSN development, e.g. standardization of communication protocols; the lack of energy-efficient power sources; the development of new ultra-low-power microelectronic components; etc.

An integration of WSN within the public data networks as well as within the domains where confidential and private data are processed (e.g. E-Health) brings along problems related to the ethical and legal questions too. Therefore, the terms as social safety or ethical safety should not be neglected.

The problematic of WSN is one of actual activities getting to the fore in the European Research Area since the issue of sensor networks was covered through "IoT" in FP7 program and strong continual extension is planned to be included also in Horizon 2020 program, especially in sections such Smart Transport; Health; Climate Action covered under Societal Challenges Pillar.

It is therefore essential to create an experience-sharing platform for scientific researchers and experts from research institutes, SMEs and companies who work in WSN domain where they can exchange some relevant skills and experiences as well as discuss upcoming trends and new ideas from this field. Moreover, the conference should also serve a function of a kind of networking platform facilitating interconnectivity between participants in case of a future collaboration.

Topics

Original contributions, not currently under review to another journal or conference, are solicited in relevant areas including, but not limited to, the following:

Development of sensor nodes and networks

- Sensor Circuits and Sensor devices – HW
- Applications and Programming of Sensor Network – SW
- Architectures, Protocols and Algorithms of Sensor Network
- Modeling and Simulation of WSN behavior
- Operating systems

Problems dealt in the process of WSN development

- Distributed data processing
- Communication/Standardization of communication protocols
- Time synchronization of sensor network components
- Distribution and auto-localization of sensor network components
- WSN life-time/energy requirements/energy harvesting
- Reliability, Services, QoS and Fault Tolerance in Sensor Networks
- Security and Monitoring of Sensor Networks
- Legal and ethical aspects related to the integration of sensor networks

Applications of WSN

- Military
- Health-care
- Environment monitoring
- Transportation & Infrastructure
- Precision agriculture
- Industry application
- Security systems and Surveillance
- Home automation
- Entertainment – integration of WSN into the social networks
- Other interesting applications

Event Chairs

Hodoň, Michal, University of Žilina, Slovakia
Kapitulík, Ján, University of Žilina, Slovakia
Micek, Juraj, University of Žilina, Slovakia
Ševcik, Peter, University of Žilina, Slovakia

Program Committee

Al-Anbuky, Adnan, Auckland University of Technology, New Zealand
Baranov, Alexander, Russian State University of Aviation Technology, Russia
Brida, Peter, University of Zilina, Slovakia
Dadarlat, Vasile-Teodor, Univiversita Tehnica Cluj-Napoca, Romania
Diviš, Zdenek, VŠB-TU Ostrava, Czech Republic
Elmahdy, Hesham N., Cairo University, Egypt
Fortino, Giancarlo, Università della Calabria

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 167–172

DOI: 10.15439/2015F368
ACSIS, Vol. 6

Smart Decision Fog Computing Layer in Energy-Efficient Multi-hop Temperature Monitoring System using Wireless Sensor Network

Krzysztof Daniluk
Faculty of Electronics and Information Technology,
Warsaw University of Technology System Control Division,
Complex Systems Group, Warsaw, Poland K.Daniluk@elka.pw.edu.pl

***Abstract*—Smart decision layer, as Fog Computing layer in WSN is presented and discussed. Working WSN sleep state mode analysis is presented in multi-hop temperature monitoring system using wireless sensor network in environmental monitoring scenario.**

The monitoring system consists of wireless sensor nodes, called motes. In laboratory experiment are used SunSPOT motes, invented by Sun Microsystems, now manufactured by Oracle. Temperature analyzing application is presented, forwarding measured data in a multi-hop way to the network host. Desktop computer plays here a role of temperature measurements database. Link-Quality Routing Protocol (LQRP) is used, which is based on Ad-hoc On-Demand Distance Vector (AODV), which is reactive routing protocol.

All gathered measurements, all data are sent to the base-station node, which is a special WSN node, directly connected via USB to the computer. Experiments were conducted taking into consideration especially SunSPOTs' processor board parameters - sleep state modes. Their processor board is turned into different sleep modes for different periods of time. A comparison of sleep mode time periods in total time of running an application is discussed. Periods of sleep states are different depending on how long is turned on the utility thread sleep.

The experiment's results show, that there is a need to build smart decision layer as fog computing layer for WSN. It should be done in order to manage in a better way environmental monitoring with usage of different kind of sensors, especially to better manage transferring big portions of data from sensors to the Cloud data storage. Sensors, which build WSN, should be distributed over specific areas, as well as fog computing devices should be able to communicate with sensors in order to transfer only relevant data to the Cloud.

Index Terms—Wireless Sensor Networks, Internet of Things, Fog Computing, Intelligent layer

I. Introduction

Experimental version of environmental monitoring system is discussed and presented as working in real-life temperature monitoring-oriented scenario using Wireless Sensor Networks (WSNs). The idea is to present application, measuring and sending gathered data in a multi-hop manner to the base-station node, connected to the computer. As a result of this paper is presented Fog Computing layer, enabling in more energy-efficient way to manage of data from WSN to the Cloud data storage. In chapter II Wireless Sensor Networks are introduced. In chapter III laboratory equipment is described. In chapter IV SunSPOT processor board energy states, experiment and results are presented. Chapter V presents, as a solution to solve experiment's problems, Fog Computing in Wireless Sensor Networks. Chapter VI summarizes the article.

II. Wireless Sensor Network

Wireless Sensor Network is a distributed system [1] with hundreds of small-size embedded devices, which are deployed densely over a specific area. Each sensor node participates in transferring data to other sensor nodes or base-stations, which are within its range. WSNs have been identified [2] as one of the most important technologies of this century. They have CPU power, radio transceiver and sensing capabilities, i.e. plenty of sensor devices can be deployed in a sensing area [5]. Such instruments can be deployed in a sensing area, where traditional networks are not able to be used. The problem is, that wireless sensor nodes are often small battery-fed devices, which underlines the fact, that their power source is very limited [2, 3]. An important issue is that the quality of wireless transmission depends on numerous external factors [6, 7, 8], like weather conditions or landform features.

In fig. 1 we see an example of wireless sensor network, with free-range motes/intermediate motes (without the black cover) and base-station on the right-hand side.

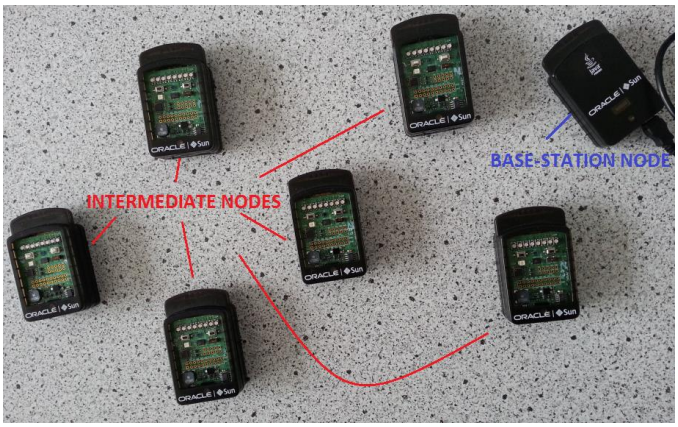


Fig. 1 Wireless sensor network structure

III. LABORATORY EQUIPMENT – HARWARDE & SOFTWARE

The laboratory equipment consists of 18 wireless sensor nodes: 12 free-range motes and 6 base-stations. Wireless sensor nodes are called SunSPOTs, they are originally invented by Sun Microsystems, now are manufactured by Oracle. The main difference between free-range mote and base-station is the fact, that free-range mote has sensor board and battery. Base-station can receive radio transmission only, when is plugged by USB to computer, no sensing capability is provided.

Figure 2 presents technical specification, of used in experiment, wireless sensor node - SunSPOT

- IEEE 802.15.4 standard for networking
- Using Squawk Java Virtual Machine
- Processor board: ARM architecture 32 bit CPU with ARM926EJ-S core running at 400MHz
- 1Mbytes SRAM memory
- 8Mbytes Flash Memory
- 2.4 GHz IEEE 802.15.4 radio with integrated antenna
- 3-axis accelerometer
- Sensors: temperature & light
- 8 tri-color LEDs, analog & digital inputs, 2 switches
- 720mAh lithium-ion battery
- Deep sleep mode: 33 uA, i.e. 909 days

Fig. 2 SunSPOT technical data

In fig. 3 we can see overview of SunSPOT's hardware, described in fig. 2

Fig. 3 Sunspot's hardware

IV. EXPERIMENT

In this part are presented following topics: A – Routing table for used multi-hop routing protocol, B – SunSPOT's processor board energy states, C – The results of experiment.

A. *Table of routing*

In the experiment LQRP routing protocol is used. This kind of protocol, offered by SunSPOT API, is based on AODV multi-hop and reactive routing protocol.

Fig. 4 presents an example of routing table, where is shown destination address, next hop address and number of hops to achieve the destination.

Destination	Next Hop	Hops
0014.4F01.0000.7C68	0014.4F01.0000.7CAC	3
0014.4F01.0000.7F92	0014.4F01.0000.7CAC	3
0014.4F01.0000.7672	0014.4F01.0000.7CAC	2
0014.4F01.0000.7631	0014.4F01.0000.7CAC	4
0014.4F01.0000.7CAC	0014.4F01.0000.7CAC	1

Fig. 4 Example of SunSPOT's routing table

B. SunSPOT's processor board energy states

Table 1 presents SunSPOT's processor board energy states [4]. This helps to understand how WSN nodes may behave, which kind of energy state may choose, depending on the situation, whether radio and sensor board are used or not. In the table 1 we can see different energy states and their corresponding radio and sensor board status, and finally current draw in particular energy states.

ENERGY STATE	RADIO status	SENSOR BOARD status	CURRENT DRAW
Deep sleep	0	0	33 µA
Shallow sleep	0	0	24 mA
Shallow sleep	1	0	40 mA
Awake, active calculating	0	0	80 mA
Awake, active calculating	1	0	98 mA
Shallow sleep	0	1	31 mA
Shallow sleep	1	1	46 mA
Awake, active calculating	0	1	86 mA
Awake, active calculating	1	1	104 mA

Table 1 SunSPOT's Processor board energy states [4] (0 = OFF, 1 = ON)

C. Results of experiments

In this section we can see experiment's results.

There is defined term **MBST – Main Board Sleep Time** calculated in milliseconds. When sleeping thread is run, it stops execution the application. The aim of the application is to forward temperature readings from one mote to another (based on routing table) in a multi-hop way. There are defined 2 types of WSN nodes: intermediate (forwarding temperature measurements to the next hop node, based on routing table) and base-station node – which collects all data from the WSN. Points of execution of the program are following:

- Inside the intermediate node, step 1 – looking for a multihop route to the base-station and opening the socket connection
- Inside the intermediate node, step 2 – sending the temperature measurement to base-station node based on routing table next-hop parameter
- Base-station node is not analyzed from the processor board energy states point of view, because is assumed it is all the time connected to the computer via USB cable

Energy state analysis covers intermediate WSN nodes.

Table 2 presents results of experiment.

	Total time of experiment (seconds)		
	10	**30**	**60**
MBST (milliseconds)	**(Shallow sleep time / total time) %**		
1	42,84%	46,64%	48,2%
2	43,61%	48,22%	50,07%
3	45,25%	48,51%	50,29%
10	45,83%	49,35%	56,57%
100	59,19%	64,55%	68,92%
1000	74,75%	86,67%	90,69%

Table 2 Experiment's results

Application, deployed on each intermediate node, is monitoring and sending by multi-hop temperature readings. **Main sleeping thread turns into sleep state for time between 1 and 1000 milliseconds.** This has influence for introducing shallow sleep time energy state, which has percentage part in total time of execution of an application.

Shallow sleep state mode of SunSPOT's processor board, is turned on, when main sleeping thread lasts less than 3000 milliseconds. After such period of idle time, deep sleep is turned on automatically, which has influence in waking up procedure all processor board elements. **From point of view of procedure of fast waking-up time, flexible change between sleep and wake-up supports shallow sleep, which seems to be the most accurate, comparing to deep sleep mode** [4].

Table 2 presents results of experiment in 3 different scenarios, when total time of experiment lasts 10, 30 or 60 seconds. In each scenario, there is run main sleeping thread, called MBST – Main Board Sleep Time, which duration is 1,2,3,10, 100 or 1000 milliseconds. For each use case is presented percentage contribution of shallow sleep time (with radio and sensor board turned on or turned off, but with no active calculations provided) in total time of running temperature analyzing application. In shallow sleep time, current draw, varies between 24mA to 46mA, depending on the fact, whether radio/sensor board is turned on or off.

D. Conclusions from the experiment

Experiment's results, shown in table 2, underline the fact, that it is hard to combine very frequent data readings (short main sleeping threads, i.e. short duration of MBST) with long lasting shallow sleep time of the whole working system.
If shallow sleep time is often, we can say, that such system is energy-efficient, i.e. is often in a sleep state mode and is able to wake up quickly. For vineyard monitoring scenario it may be crucial, for different weather conditions, to be ready for very changeable, in frequency of gathering, data readings. This happens also for data readings taken very rarely, if no major changes appear in weather conditions.
In order to work in that way, i.e. flexible changing frequency of gathering data from sensors, dependable on weather conditions, as well as better managing big data taken from WSN, Fog Computing layer is introduced to use in WSN.

V. FOG COMPUTING IN WSN

A. Fog Computing in WSN

In this chapter is proposed Fog Computing layer for WSNs in order to manipulate in more efficient way [9, 10] with different energy states in SunSPOTs (shallow sleep, deep sleep, awake).
Fog Computing layer should be understood as a bridge between the Cloud data center (where WSN data readings may be stored) and the Internet of Things, here WSNs.
Fog is a distributed computing architecture, a paradigm extending Cloud Computing to the edge of the network. Fog, like Cloud, provides data, computation, storage and application services to end-users. The main characteristics for the Fog is its proximity to end-users, used especially to reduce service latency, enabling real time big data analytics.
Fog devices are heterogeneous devices like access points, edge routers and switches, end-user devices. The Fog platform supports real-time analytics processes, filters the data and pushes to the Cloud. There is waste of network resources to transmit plenty of gigabytes of data to the cloud and also in doing analytics in that manner. Better way is to analyze data locally and then decide what should be passed to the Cloud and what requires immediate reaction.
There is presented architecture for Fog Computing layer, with gathering data from WSN, analyzing and forwarding only the necessary data for the end user. Fog Computing devices, in their analytical work, should cover such parameters like: packet size, signal strength, volume traffic controller, radio controller, sensor board controller & master clock controller of each sensor from the WSN.
The aim of Fog Computing smart decision layer in WSN is to control shallow and deep sleep mode in more flexible way [19, 20], as well as to better manage Big Data, i.e. doing the whole analytical work.
Nowadays, whole analytical work is done in WSN by each sensor separately. In this architecture, whole analytical work will be done by Fog Computing devices, which collect sensor readings and give them feedback with information to change e.g. their frequency of the readings.
Fog Computing devices may also act, as control devices of Wireless Sensor Networks, i.e. they may store in their databases the battery level of each sensor, calculate the battery level of each sensor based on transmission analysis, without asking each sensor about its battery level, i.e. without asking about additional data from sensors, which may be calculated by Fog Computing layer. **The next step for further experiments is to start to build Fog Computing layer for WSN in order to enhance the energy-efficient work [13, 14, 15] of wireless sensor network nodes not only based on turning on/off the main sleep scheduler, like done in the experiment, but also with manipulating the work [12] of master clock controller, changing the signal strength, introducing volume traffic controller in Fog Computing layer.**

B. Future work

Next experiment will introduce decreasing the computational power, controlling the traffic volume and radio signal strength. It is assumed [11], that these new parameters will extend the length of shallow sleep duration, so will introduce new energy-efficient way [16, 17, 18] for resource scheduling in Wireless Sensor Networks. Table 3 presents concept for new parameters in Fog Computing layer in WSN.

WSN Fog Computing layer (Resource scheduling of each sensor from WSN)	
PARAMETERS	**OBJECTIVE FUNCTION**
1) Packet size	Shortest data frame with deep sleep indication flag
2) Master Clock Controller	Decreasing the computational power after load level analyzing
3) Signal strength (transmission power control)	Decreasing the radio power, if received signal strength indicator (RSSI) give assumption it is safe and possible
4) Radio Controller	Radio on/off, radio channel
5) Sensor Board Controller	Sensors on/off connected
6) Volume traffic controller	Radio Controller + Sensor Board Controller

Table 3 Concept for new parameters in WSN Fog Computing layer

Fig. 5 presents new concept of Fog Computing layer in WSN, which will be introduced in experiments analyzing duration of shallow-sleep time in total time of working application in WSN.

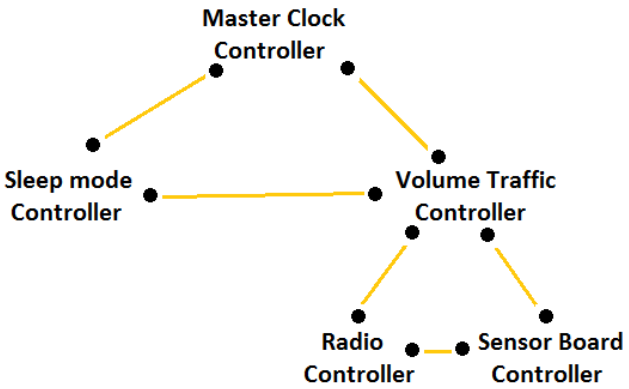


Fig. 6 Concept of Fog Computing layer in WSN

Below are proposed main parts of Fog Computing layer in WSN:

- **Master Clock Controller**

  Analysis and changing master clock frequency

- **Sleep Mode Controller**

  Analysis and changing sleep mode (deep, shallow, active)

- **Volume Traffic Controller**

  Analysis how much data where transferred in amount of time in the past – analysis for the future, expected traffic

- **Radio Controller**

  Analysis and turning on/off the radio. Changing radio power (RSSI)

- **Sensor Board Controller**

  Analysis and turning on/off selected sensor (temperature, light, accelerometer)

VI. Conclusion

Fog Computing layer in WSN is presented as a solution for problems presented in the experiment. Such architecture layer should be implemented in order to better manage vineyard monitoring transmissions for different kind of sensors and especially to better manage transfer of Big Data from sensors to the Cloud data storage. Fog Computing devices should be able to communicate with sensors in order to transfer only relevant data to the Cloud. Concept of Fog Computing layer in WSN is presented and as a future work is going to be implemented in the next experiment. Presented concept supports energy-efficiency in Wireless Sensor Networks.

References

[1] M. C. Vuran, I. F. Akylidiz, Wireless Sensor Networks. Wiley, 2010.

[2] A. Tiwari, P. Ballal, and F. L. Lewis, Energy-efficient wireless sensor network design and implementation for condition-based maintenance, ACM Trans. Sensor Networks (TOSN), vol. 3, no. 1, pp. 1–23, 2007.

[3] S. R. Gandham, M. Dawande, R. Prakash, and S. Venkatesan, S., Energy efficient schemes for wireless sensor networks with multiple mobile base stations, in Proc. IEEE Global Telecom. Conf. GLOBECOM'03, San Francisco, USA, 2003, vol. 1, pp. 377–381.

[4]https://java.net/projects/spots-core-libraries/sources/svn/content/trunk/sdkresources/resources/doc/SunSPOT-Programmers-Manual.pdf (accessed: May 2015)

[5] P. Baronti, P. Pillai, V. W. C. Chook, S. Chessa, A. Gotta, and Y. Fun Hu, "Wireless sensor networks: a survey on the state of the art and the 802.15.4 and ZigBee standards", Comp. Commun., vol. 30, no. 7, pp. 1655–1695, 2007.

[6] ZigBee Alliance, "ZigBee Specification v1.0", New York, USA, 2005.

[7] W. R. Heinzelman, A. Chandrakasan, H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks", in Proc. 33rd Annual Hawaii Int. Conf. Sys. Sciences HICSS'00, Maui, Hawaii, USA, 2000, pp. 3005–3014.

[8] B. Azzedine, C. Xiuzhen, and J. Linus, "Energy-aware datacentric routing in microsensor networks", in Proc. 6th Int. Symp. Model. Analys. Simul. Wirel. Mobile Sys. MSWiM 2003, San Diego, CA, USA, 2003, pp. 42–49.

[9] K. Lin, Ch. F. Lai, X. Liu, and X. Guan, "Energy efficiency routing with node compromised resistance in wireless sensor networks", Mob. Netw. Appl., vol. 17, pp. 75–89, 2012.

[10] M. I. Shukur, L. S. Chyan, and V. V. Yap, "Wireless sensor networks: delay guarantee and energy efficient MAC protocols", World Academy of Sci., Engin. Technol., vol. 50, pp. 1061–1065, 2009.

[11] S. R. Gandham, M. Dawande, R. Prakash, and S. Venkatesan, S., Energy efficient schemes for wireless sensor networks with multiple mobile base stations, in Proc. IEEE Global Telecom. Conf. GLOBECOM'03, San Francisco, USA, 2003, vol. 1, pp. 377

[12] K. Daniluk, E. Niewiadomska-Szynkiewicz, Energy-Efficient Security in Implantable Medical Devices, FedCSIS 2012 Proceedings,pp.773-778

[13] H. Cam, S. Ozdemir, D. Muthuavinashiappan, and P. Nair, Energy efficient security protocol for wireless sensor networks", in Proc. IEEE 58th Veh. Technol. Conf. VTC 2003, Orlando, Florida, USA, 2003, vol. 5, pp. 2981–2984.

[14] Zhu, S. Setia, and S. Jajodia, "LEAP+: Efficient security mechanisms for large-scale distributed sensor networks", ACM Trans. Sensor Netw. TOSN, vol. 2, no. 4, pp. 500–528, 2006

[15] Zhu, S. Setia, and S. Jajodia, "LEAP: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks", in Proc. 10th ACM Conf. Comp. Commun. Secur. CCS 2003, Washington, DC, USA, 2003, pp. 62–72.

[16] L. E. Lighfoot, J. Ren, and T. Li, "An energy efficient link-layer security protocol for wireless sensor networks", in Proc. IEEE Int. Con. Elec.-Infor. Technol. EIT 2007, Chicago, IL, USA, 2007, pp. 233–238.

[17] S. K. Singh, M. P. Singh, and D. K. Singh, "Energy-efficient homogenous clustering algorithm for wireless sensor networks", Int. J. Wirel. Mob. Netw., vol. 2, no. 3, pp. 49–61, 2010.

[18] S. K. Singh, M. P. Singh, and D. K. Singh, “A survey of energefficient hierarchical cluster-based routing in wireless sensor networks”, Int. J. Adv. Netw. Appl., vol. 2, no, 2, pp. 570–580, 2010.
[19] C. Castelluccia, A. C.-F. Chan, E. Mykletun, and G. Tsudi k, “Efficient and provably secure aggregation of encrypted data in wireless sensor networks”, J. ACM Trans. Sensor Netw. (TOSN), vol. 5, no. 3, 2009
[20] Niewiadomska-Szynkiewicz, P. Kwaśniewski, and I. Windyga, “Comparative study of wireless sensor networks energy-efficient topologies and power save protocols”, J. Telecom. Inform. Technol., no. 3, pp. 68–75, 2009

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems. Events that constitute IT4BMS are:

- **ABICT'15** - 6th International Workshop on Advances in Business ICT
- **AITM'15** - 13th Conference on Advanced Information Technologies for Management
- **ISM'15** - 10th Conference on Information Systems Management
- **IT4L'15** - 4th Workshop on Information Technologies for Logistics
- **KAM'15** - 21st Conference on Knowledge Acquisition and Management

6th International Workshop on Advances in Business ICT

ABICT focuses on Advances in Business ICT approached from a multidisciplinary perspective. It will provide an international forum for scientists/experts from academia and industry to discuss and exchange current results, applications, new ideas of ongoing research and experience on all aspects of Business Intelligence. ABICT will be also an opportunity to demonstrate different ideas and tools for developing and supporting organizational creativity, as well as advances in decision support systems.

We kindly invite contributions originating from any area of computer science, information technology and computational solutions for different applications areas, data integration and organizational implementation of ABICT, as well as practical ABICT solutions.

Topics

Topics include (but are not limited to):

- Advanced technologies of data processing, content processing and information indexing
- Analytics as a service
- Big Data: benefits and challenges
- Business Analytics
- Business applications of social networks
- Business data mining and knowledge discovery
- Business Intelligence
- Business Rules
- Business-oriented time series data mining, analysis, and processing
- Cloud based Business Intelligence
- Creativity Support Tools
- Customer Relationship Management, social Customer Relationship Management
- Data driven marketing
- Data Warehousing
- Decision support
- Digital Business Strategy
- Enterprise Device Management
- ICT technologies in enterprise management
- Information forensics and security, information management, risk assessment and analysis
- Information Systems Design
- Internet of Things
- Knowledge Management (for better Decision Support, Collaboration and Competitiveness)
- Legal text processing
- Leveraging ICT for Transforming Organization
- M2M Device Management, M2M Solutions
- Semantic Web and Ontologies in Business ICT
- Virtual Enterprise
- Web 2.0 and Web 3.0 in fusing Business Intelligence systems and Decision Support Systems
- Web-Based Data Management Systems

Event Chairs

Mach-Król, Maria, University of Economics in Katowice, Poland

Olszak, Celina M., University of Economics in Katowice, Poland

Pełech-Pilichowski, Tomasz, AGH University of Science and Technology, Poland

Program Committee

Abramowicz, Witold, Poznan University of Economics, Poland

Badica, Amelia, University of Craiova, Romania

Berio, Giuseppe, Universite de Bretagne Sud, France

Chiu, Dickson K. W., Dickson Computer Systems, Hong Kong S.A.R., China

Christozov, Dimitar, American University in Bulgaria, Bulgaria

Gaweł, Bartłomiej, AGH University of Science and Technology

Kacprzyk, Janusz, Institute of Computer Science, Polish Academy of Sciences, Poland

Khachidze, Manana, Tbilisi State University, Georgia

Konikowska, Beata, Institute of Computer Science, Poland

Korwin-Pawlowski, Michael L., Universite du Quebec en Outaouais, Canada

Kulczycki, Piotr, Systems Research Institute, Polish Academy of Sciences, Poland

Loucopoulos, Peri, Harokopio University of Athens, Greece

Madeyski, Lech, Wrocław University of Technology

Ogihara, Mitsunori, University of Miami, United States

Owoc, Mieczyslaw, Wroclaw University of Economics, Poland

Petryshyn, Lubomyr, AGH University of Science and Technology, Poland

Prasad, T. V., Visvodaya Technical Academy, India

Pulvermueller, Elke, University Osnabrueck, Germany

Reimer, Ulrich, University of Applied Sciences St. Gallen, Switzerland

Rossi, Gustavo, National University of La Plata, Argentina

Salem, Abdel-Badeeh M., Ain Shams University, Egypt

Sauer, Jurgen, University of Oldenburg, Germany

Szpyrka, Marcin, AGH University of Science and Technology, Poland

Teufel, Stephanie, University of Fribourg, Switzerland

Zieliński, Jerzy S.

Zurada, Jozef, College of Business University of Louisville, Louisville

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 177–184

DOI: 10.15439/2015F17
ACSIS, Vol. 6

Cognitum Ontorion: Knowledge Representation and Reasoning System

Paweł Kapłański
Gdansk University of Technology
Department of Applied Informatics in Management
Faculty of Management and Economics
Gabriela Narutowicza 11/12, 80-233 Gdansk, Poland
Email: pawel.kaplanski@zie.pg.gda.pl

Paweł Weichbroth
Gdansk University of Technology
Department of Applied Informatics in Management
Faculty of Management and Economics
Gabriela Narutowicza 11/12, 80-233 Gdansk, Poland
Email: pawel.weichbroth@zie.pg.gda.pl

***Abstract*—"If knowledge can create problems, it is not through ignorance that we can solve them." (Isaac Asimov). Nevertheless, at any point of human activity, knowledge (besides practice) is a key factor in understanding and solving any given problem. Nowadays, computer systems have the ability to support their users in an efficient and reliable way. In this paper we present and describe the functionality of the Cognitum Ontorion system. Firstly, we identify emerging issues focused on knowledge representation and reasoning. Secondly, we briefly discuss models and methodology of agent-oriented analysis and design. Next, the semantic knowledge management framework of the system is reviewed. Finally, the usability of Ontorion is argued based on a case study, in which a software process simulation modeling environment is developed. At the end we provide future work directions and final conclusions.**

I. Introduction

DAVID GARVIN notes that "to move ahead, one must often first look behind" [1]. Let us ask this cinch question: how do you want to understand the past, sense the present and predict the future, if you are unable to preserve your knowledge? Indeed, a lot of human effort and material resources have been exploited in preserving knowledge. Indubitably, humans have been using many different forms to express and share knowledge (e.g. drawings, symbols, words and numbers, which nowadays are commonly encoded in computer memory). Now, let us focus on some formal representation methods of knowledge which are the prominent subfield of artificial intelligence (AI).

In the AI canon, knowledge seems to be always defined in a strictly functional way. However, from all incoming questions to the mind of an attentive reader, which one would be the first to reveal the question: What is knowledge? – This might be the question for the majority of us. Bearing in mind a computer "brain" – central processing unit (CPU) is principally able to process sequences of bits where a single bit is represented by two exclusive numbers: 0 (zero) or 1 (one), as a consequence, at the moment we are able to represent "only" facts and rules in computer memory. A distinct fact (or a set of facts), represented by a sentence (or a set of sentences), is used in deductive reasoning. A single rule (or a set of rules), expressed in a form: $if \rightarrow then$, may be a logic or be inductive in its genesis. Elements of particular knowledge (intra- or inter-connected facts or rules) are often named as "knowledge chunks" [2]. A single chunk is commonly attached to an exclusive agent (an independent and separate application unit).

In the beginning, AI research investigated how a single agent can exhibit singular and internal intelligence. However, in recent years, we have observed an interest in concurrency and distribution in AI which have been named as distribution artificial intelligence (DAI). This recent discipline can be divided into two primary areas: distributed problem solving (DPS) and Multi-Agent (MA) systems. It is not a straightforward task to coordinate knowledge, goals and actions among a collection of autonomous agents.

Some successful application of agent-oriented architecture can be pointed in decision support systems (DSS) in the area of the discovery of stock market gamblers patterns [3], [4], web usage mining [5] and the evaluation of information technology [6].

II. Reasoning in agent-oriented design and analysis. A hybrid approach

There is a long history of symbolic reasoning usage in order to provide intelligent behavior in Multi-Agent & Simulation systems (MASS). Deductive Reasoning Agents, which use logic to encode a theory defining the best action to perform in a given situation, are the "purest" in terms of their formal specification. Unfortunately, they suffer from all the practical limitations of formal representation: firstly, the complexity of theorem proofs (it may even lead to undecidable statements) and secondly, the boundaries of expressivity formed by core knowledge representation attributes (e.g. monotonicity of knowledge, open world assumption).

Making deductive reasoning requires the selection of underlying logics that support the nature of agents. It is worth mentioning that the most prominent implementations of deductive reasoning agents are based on intentional logics like formal models of intention logics [7] (e.g. Belief – Desire – Intention, BDI), which take into account some subset of the Saul Kripke modal logic [8].

Problems with symbolic reasoning led to the establishment of the "reactive agent movement" in 1985, revealing an era of reactive agent architecture. The reactive agent movement

manifested in the form of requirements for so-called behavior languages [9]:

1. Intelligent behavior can be generated without explicit representations of the kind that symbolic AI proposes.
2. Intelligent behavior can be generated without explicit abstract reasoning of the kind that symbolic AI proposes.
3. Intelligence is an emergent property of certain complex systems.

Reactive agents are nowadays well recognized but still they lack formal foundations and therefore these kind of MASS are very hard to analyze with formal methods and tools. Nevertheless, the reactive agent movement resulted in Agent Oriented Programming (AOP), e.g. JADE [10], which is currently considered as a step beyond Object Oriented Programming (OOP) in Software Engineering.

A novel approach to designing MA systems – Hybrid Agent Architecture, attempts to combine the best of symbolic and reactive architectures. The system itself is built up of at least two subsystems: (1) a symbolic world model that allows plans to be developed and decisions made and (2) a reactive engine which is capable of reacting to events without involving complex reasoning.

As an example let us consider Ferguson's "TouringMachine" [11], which fits into the definition given above. Ferguson defines: *"The TouringMachine agent architecture comprises three separate control layers: a reactive layer, a planning layer, and a modelling layer. The three layers are concurrently-operating, independently-motivated, and activity-producing: not only is each one independently connected to the agent's sensory apparatus and has its own internal computational mechanisms for processing appropriate aspects of the received perceptual information, but they are also individually connected to the agent's effectory apparatus to which they send, when required, appropriate motor-control and communicative action commands"*.

We present a novel approach to hybrid agent architecture, which is implemented on top of a scalable Knowledge Representation & Reasoning (KRR) system. KRR allows each environment to be described formally as well as giving the possibility to build a reactive agent system based on a knowledge base (KB) triggering subsystem. Moreover, it provides agents with synthesis tools.

Here, we consider a reactive agent that is able to maintain its state [12]. This agent has an internal stand-alone data structure, which is typically used to record information about the state and history of the environment and to store a set of all the internal states of an agent (Fig. 1).

The perception of an agent is realized in its see function if the function is time-independent. The agent's action selection is defined as a mapping from its internal states to actions. The $next$ function maps an internal state and percept to an internal state. The abstract agent control loop is then:

1. Start with the initial internal state $s \leftarrow s_0$.
2. Observe the environment state e, and generate a percept $p \leftarrow see(e)$.

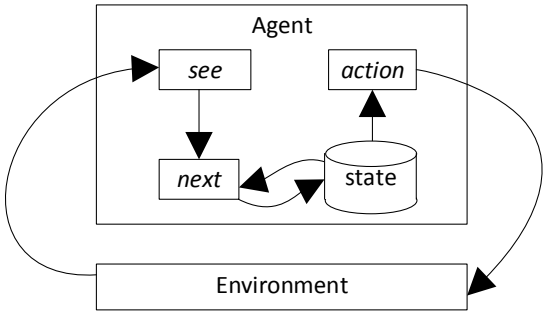


Fig. 1. An agent with its internal state

3. Update the internal state via the $next$ function $s \leftarrow next(s, p)$.
4. Select an action via the $action$ function $a \leftarrow action(s)$.
5. GOTO 2.

The environment state e is (in hybrid architecture) given by the symbolic system – here we use ontology to encode it. The function $next$ is one that needs to be implemented either by the programmer or by an automated process. In the first case, however, it is hard to distinguish such an agent from a (considerably complex) object oriented program (formerly, active-object design pattern implementation [13]). On the other hand, automated agent synthesis is an automatic programming task: given an environment, let's try to automatically generate an agent that succeeds there. The synthesis algorithm should be both sound and complete. Sound means here that the agent will succeed in the given environment once it is correctly constructed, and completeness guarantees the possibility to create the agent for the given environment.

The hybrid approach allows us to build a semi-formal foundation for MASS that allows for a sound and complete synthesis of agents as long as their definition fits into the expressivity frame of underlying logic and if the underlying logic has the reasoning task to be sound and complete itself. This is true for Description Logic (DL) [14] – the foundation for OWL [15], therefore we selected OWL compliant KRR.

III. Ontorion Architecture

Modern Scalable Knowledge Management Systems give the possibility to use KRR in a similar way as we tend to use RDBMS. We have focused on a KRR system the functionality of which allows a user-interface in natural language to be implemented and used.

Ontorion [16] is a Distributed Knowledge Management System that allows semi-natural language to be used to specify and query the knowledge base. It also has a built-in engine trigger which fires the rules each time if the corresponding knowledge is modified. Ontorion supports the major W3C Semantic Web standards: OWL2, SWRL, RDF, SPARQL. Ontologies can easily be imported from various formats, exported to various formats, and accessed with SPARQL [17]. Solutions built on

top of Ontorion can be hosted both in the Cloud and On-Premise environments.

By design, Ontorion allows one to build large, scalable solutions for Semantic Web. The scalability is realized by both – the noSQL, symmetric database Cassandra [18] and the internal ontology modularization algorithm [19]. Ontorion is a cluster of symmetric Nodes, able to perform reasoning on large ontologies. Every single system node is able to do the same operations simultaneously on data sets – it tries to get the minimal suitable ontology module (part) and perform any requested task on it.

The symmetry of the architecture of the cluster provides system scalability and flexibility – Ontorion can be deployed and executed in a computing cloud environment, where the total number of nodes can be changed on request, depending on user requirements.

The fundamental algorithm in a KRR system such as Ontorion ought to reason over description logic selected as a foundation for OWL called $\mathcal{SROIQ}^{(\mathcal{D})}$ [20], and should be able to process complete or selected segments of ontologies.

If performance is more important than expressivity power, then it is possible to switch Ontorion into OWL-RL+ mode. OWL-RL+ mode is constructed in a similar way to how it was first implemented in DLEJena [21]. The reasoning process in OWL-RL+ mode remains in $\mathcal{SROIQ}^{(\mathcal{D})}$ for the T-Box, while for the A-Box the reasoning is based on the OWL-RL ruleset.

Furthermore, the modular separation of complex ontologies also allows the reasoning process to be partitioned, which can be performed on knowledge modules (independent pieces of knowledge) – in parallel, at the same time on separate machines. In other words, the modularization algorithm is scalable and traceable.

In Ontorion, conclusions that are the results of new incoming knowledge can fire triggers at extension/reactive points (Fig. 2). On the other hand, if some chunk of knowledge meets a set of predefined conditions, a knowledge modification trigger executes the procedures responsible for interaction with external systems (e.g. sending a notification using an SMTP server). We observed that knowledge modification triggers allow the Hybrid Agent MASS to be built on top of the Ontorion KRR.

The underlying storage for Ontorion is the BigTable [22] implementation (namely Cassandra), which is able to maintain a petabyte of data. Together with an analytic cluster e.g. Hadoop [23] it forms a BigData solution. In these terms, we can consider Ontorion as a BigKnowledge solution and our Agent application as a BigAgent, which together constitutes a highly scalable hybrid agent infrastructure.

Distributed systems lack the one common model of time. In a distributed environment, time is relative and the serialization of events requires an internode negotiation algorithm. In modern distributed noSQL databases like Cassandra 2.0, the serial time model can be preserved with the use of the Paxos algorithm [24], which allows a distributed atomic "Compare and Set" (CAS) functionally to be efficiently implemented.

CAS in Ontorion is used to maintain the Agent state in a coherent way, therefore the existence of CAS is critical for proper system functioning. CAS also provides a way to make the Agent System fault-tolerant by transactional-queue implementation, which is crucial for long-running simulations.

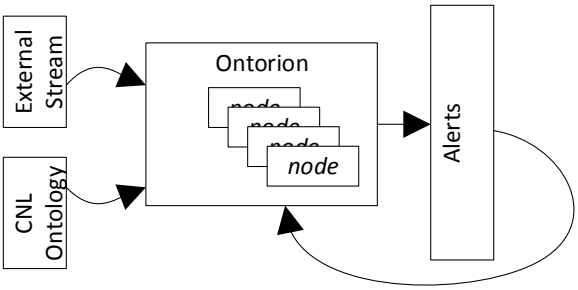


Fig. 2. The Ontorion KRR

IV. Programming Agents with Natural Language

An important, novel feature of Ontorion, among other KRR systems, is the ability to describe knowledge and interact with the user in semi-natural language. The language represents the family of controlled natural languages that is expressive enough to describe OWL. Controlled natural language (CNL) is a subset of natural language with a reduced grammar and vocabulary, which – in this case – translates directly to logic with formal semantics capabilities.

In general, controlled natural language should be unambiguous and intuitive, ultimately forming an easy way for human-machine interaction (understandable by humans, executable by machines). Due to its limitations, it needs to be supported by a predictive (structural) editor which is Ontorion FluentEditor tool [25]. The other, well-known implementation of a controlled natural language is "Attempto Controlled English (ACE)" [26], developed by the University of Zurich. However, the origins of CNL can be found in the famous novel by George Orwell: "1984", where he discusses the NEWSPEAK – a controlled language. The most used industrial implementations nowadays are Domain Specific Language (DSL) (implemented as a part of the Drools project) [27] and Semantics of Business Vocabulary and Rules (SBVR) [28], whereas CNL allows the representation of BPML diagrams.

In Ontorion Fluent Editor controlled language is equipped with formal semantics expressed in logic. General groups of sentences are allowed which include:

1. Concept subsumption, represents all cases where there is a need to specify (or constrain) the fact about a specific concept or instance (or expressions that evaluate the concept or instance) in the form of subsumption (e.g.: *Every cat is a mammal*, *Pawel has two legs* or *One cat that is a brown-one has red eyes*).

2. Role (possibly complex) inclusion specifies the properties and relationships between roles in terms of the expressiveness of $\mathcal{SROIQ}^{(\mathcal{D})}$ (e.g.: *If X loves something that covers Y then X loves-cover-of Y*).
3. Complex rules; If [body] then [head] expressions that are restricted to the DL-Safe SWRL subset [29] of rules (e.g.: *If a scrum-master is-mapped-to a provider and the scrum-master has-streamlining-assessment-processes-sprints-level equal-to 2 then the provider has-service-delivery-level equal-to 1 and the provider has-support-services-level equal-to 2*).
4. Complex OWL expressions; the grammar allows the use of parentheses that can be nested if needed in the form of (that) e.g.: *Every human is something (that is a man or a woman or a hermaphrodite).*
5. Aforementioned knowledge modification triggers that have the form of: *If P then for-each P execute Q*, where P is a premise and Q a consequence. Premise P is an expression that evaluates a set of connected instances that fulfill some conditions, while the consequence Q is a procedure written in $C\#$ programming language (e.g.: fig. 5).

V. The definition of the Multi-Agent system in terms of Knowledge Management System triggers

Here, we present a modern scalable KRR as a foundation for MASS. The discussed KRR (Ontorion) enables the specification of knowledge-modification triggers in the form of reactive rules: if $\rightarrow$ action. Ontorion knowledge modification triggers allow the knowledge itself to be modified and therefore it is possible to build a set of triggers here that are fired continuously. A reactive trigger like this breaks the decidability of the underlying knowledge base and, as a consequence, KRR tasks based on Ontorion are decidable only if all deductive rules are DL-Safe (e.g. they are SWRL equivalent) – otherwise these tasks are non-decidable.

The above property of system modification triggers is very useful for the development of the hybrid MASS. The hybrid agent paradigm can be adapted, by using triggers, even if the environment is modelled in Ontorion as an OWL Ontology, with all its limitations (e.g.: lack of modality or time representation). We can define agents here as OWL individuals. The behaviour of the agents is implemented in reactive rules called $moves$. These rules combine the see, $next$ and $action$ functions discussed earlier (the reactive, state based and abstract model of an agent). Moreover, agent-individuals and "ordinary" OWL individuals are different as agent-individuals are equipped with a transactional, CAS protected internal state, represented by related data-values.

A single $move$ function as a parameter takes a percept which is a result of a reasoning process (here, we consider reasoning as an implementation of the abstract see function) over the current state of the environment. In the implementation of this function a CAS operation is used to preserve transactional semantics. The $move$ function is only activated if the perceived-message is equal to the expected-message.

```
If ... then ... execute <?
    Move(agent, "current-state", message,
        "expected-message", ()=>
    {
         /* the agent action */
        return "new-state";
    });
?>.
```

Fig. 3. General form of the $move$ function

There is not one single agent that activates on perceived-message - it can be any agent that fulfills the rule premise, therefore a rule conclusion can be reused by many agents and the overall execution result of the system is non-deterministic. Messages are transferred between agents using a distributed message queuing system, managed by the KRR.

A single agent is determined by its state and all the $move$ functions that can be ever executed in the context of its state, therefore here, the agent synthesis process, is a process of the assignment of the $move$ functions to the single agent-individual.

The reactive-rule bodies of the $move$ function determine the specific environment state that allows the system to assign the function to the agent; however, the overall behaviour of MASS is non-deterministic. This is due to the fact that the concrete run (the MASS run) needs the selection of agent instances made in runtime – and runtime (in opposition to reasoning time) is a part of the reactive model that is non-deterministic by nature. The non-deterministic selection of choices, often by use of pseudo-random number generators, and the parallel execution of different threads, are required by underlying technologies to provide an efficient computational model.

Therefore, we need to keep in mind, that simulations based on the reactive/hybrid approach are non-deterministic. In this case, we have to perform a large set of experiments with the same initial state and then use analytical tools and methods to verify the statistical hypothesis.

VI. The scalability of KRR oriented MASS

In practical scenarios, when it comes to simulating large societies, it is important to simulate a large amount of agents at the same time. From the technological perspective, currently, we can model large societies of people. Nowadays, the existence of 7×10^9 beings can be encoded in less than 1 GB of memory. If we encode a single human being as a 1 kB vector of bits then we can store an entire population in a single modern hard drive at a relatively low cost. Working with cloud-based environments, we can hire thousands of computers for a few hours with a similar amount of money. Therefore MASS scalability – the ability of the system to scale together with the

size of the problem - is regarded as a mandatory and critical property.

Ontorion is a scalable KRR system, approximately associated with the size of maintained knowledge due to modularization algorithms embedded, whereas Cassandra, as the underlying storage solution, is scalable by its design.

In distributed systems, task synchronization is a burden and sometimes even an obstacle. Task distribution over a set of physical machines demands synchronization protocols. Satisfyingly, the Cassandra database has the Paxos protocol implemented, which allows a global CAS functionality to be implemented. The ability of agents to modify particular chunks of knowledge indicates influence on the surrounding environment as well. What can be seen as a common task in terms of RDBMS (e.g. some database modification), might have large and complex implications in terms of a distributed knowledge base. Given the subset of First Order Logic (FOL), we deal with the monotonic knowledge model. The monotonicity implies that there is no impact on the overall meaning when the order of adding knowledge is one way or another.

When we modify knowledge the problem is somehow more complicated – besides agents tend to modify knowledge very often. The cost of knowledge modification depends on its level, scale and size. The relation between knowledge generality and its modification cost is positive as a result of the replacement of all revalued conclusions. Moreover, knowledge modification triggers, used to implement the *next* functions, break the open world assumption (OWA) [14]. This effect is caused by their ability to modify knowledge depending on the "known" parts of the knowledge. An agent may learn knowledge even when it stays in contradiction to what it already knows. In addition, knowledge modification triggers break the monotonicity of the knowledge base. Therefore, the order of agent Next firing is significant in terms of the final knowledge base shape. As previously mentioned, simulations based on the reactive/hybrid approach are non-deterministic due to both the distributed system properties and the internal non-determinism of the reactive agent system.

VII. Experimental setup - Software Process Simulation Modelling (SPSM) Environment

Software Process Simulation Modelling (SPSM) [30] is widely used nowadays to support planning and control during software development [31], [32]. MA systems play a very important role here as they naturally can be used to simulate social behaviors in the software testing phase. In our approach, the SPSM is divided into two components: ontology and knowledge modification triggers. In the example given below (see Fig. 4), the ontology defines (with CNL) the core concepts such as: competency, task, developer, manager:

We also defined agent-rules by making use of knowledge modification triggers (see Fig. 5,6). Those triggers implement the following scenario: a developer with certain competencies starts to realize a task. After the task is finished, new knowledge about the task realization process is added, and a "Busy" state is set on the developer. The second trigger is fired when the task is finished and a "Ready" state is set back.

```
Cpp-Programming is a competency.
Java-Programming is a competency.
...
Task-0 is a task.
Task-1 is a task.
Task-1 is-dependent-on Task-0.
Task-1 requires-competency Cpp-Programming.
Task-1 has-estimated-realization-md equal-to 500.
Task-2 is a task.
Task-2 is-dependent-on Task-1.
Task-2 requires-competency Java-Programming.
Task-2 has-estimated-realization-md equal-to 500.
...
Anna is a developer .
Anna has-competency Cpp-Programming .
Anna has-competency Java-Programming .

John is a developer .
John has-competency Java-Programming .
John has-competency Web-Programming .
...
```

Fig. 4. Configuration of the SPSM environment

Every time the environment contains a situation where one task is dependent on the other, finished task, we execute the trigger that forces previous triggers to start a simulation.

Here, we use the *once* function, which ensures that the execution of the trigger happens exactly once (note that this is not a trivial task, not executed in a stand-alone system but in a distributed system which requires dispersed CAS operation). The last step is to define the simulation entry point (see fig. 7).

The start event (see fig. 8) sets up the agents and defines the initial task for the "Done" state to activate the overall simulation.

In the above simulation, to make a long story short, in the beginning, we defined three distinct sets: task, competence and individual. Each agent represented a particular individual (developer). Each task required a precise competency and was time specified. This simplified description of the modelled micro-world was given as an input to the system. Next, on a user request the simulation was executed and a set of rules started to be processed due to accomplish a given set of tasks (Fig. 10).

Ontorion usability has been evidenced in one of many possible applications. In highly complex systems or projects, it seems that it is a considerable issue to design, estimate and finally test all possible dependency relationships between processes and their execution sequence. We showed how to optimize the selection of a developer's competency to a particular tasks. In this instance, we were able to identify "hidden" bottlenecks and constraints.

The experiments performed on the Ontorion cluster show

```
If a task-realization-query requires-competency a
competency and a developer has-competency the
competency and the task-realization-query has-origin
a task then for the task-realization-query and the
developer and the task execute <?
    Move(developer, "Ready",
      task_realization_query, "Programming", ()=>
    {
// read the realization time
var realizationTime =
  (from v in Values where
      v.source==InstanceDL(task) &&
      v.datarole=="have-estimated-realization-md"
   select v.value).
   FirstOrDefault().
   SetConsistencyLevel(ConsistencyLevel.Quorum).
   Execute();

// create the wake-up message
        var msgid = CreateMessage(developer,
"WakeUp",task);

// delayed (by the realization time) modification //
of KB
  KnowledgeInsertWithDelay(
      msgid + " is a wake-up-message."+
      msgid + " has-origin " + developer + "."+
      msgid + " has-task-realization-query "
            + task_realization_query + "."+
      msgid + " has-target "+ developer + ".",
   int.Parse(realizationTime));

// mark the agent state as busy
        return  "Busy";
    });
?>.
```

Fig. 5. The *move* function written in C# is fired when the developer is ready and fit for the given task

```
If a wake-up-message has-target a developer and the
wake-up-message has-origin the developer and the wake-
up-message has-task-realization-query a task-
realization-query and the task-realization-query has-
origin a task then for the wake-up-message and the
developer and the task-realization-query and the task
execute <?
    Move(developer, "Busy",
         wake_up_message, "WakeUp", ()=>
    {
// modify the status of task
       KnowledgeInsert(task+" has-status Done.");

// mark the agent state as ready
       return "Ready";
    });
?>.
```

Fig. 6. The *move* function written in C# is fired when the developer is done with the task

```
If a start-event exists then for the start-event execute
<?
    Once("Lets the simulation start.", ()=>
    {
      CreateAgent("Mark","Ready");
      CreateAgent("John","Ready");
      CreateAgent("Tom","Ready");
      CreateAgent("Gabi","Ready");
      CreateAgent("Anna","Ready");

      KnowledgeInsert("Task-0 has-status Done.");
    });
?>.
```

Fig. 7. The simulation entry point

```
Start-Event is a start-event.
```

Fig. 8. The simulation start event

flexible system scalability, persistent intra-communication duration between nodes and overall system stability. As an illustration, let us present this factual operation. A new node added to the server farm was properly initialized and broadcasted to other nodes and a scheduled job was again distributed. Still, a systematic empirical measurement needs to be made to monitor system behavior, especially when some changes have taken place. A cloud-based environment is our choice due to the obvious benefits of machine virtualization.

For Ontorion cluster setup, we used cluster of 3 standard VM nodes. On top of this cluster we executed MASS made of 5 agents.

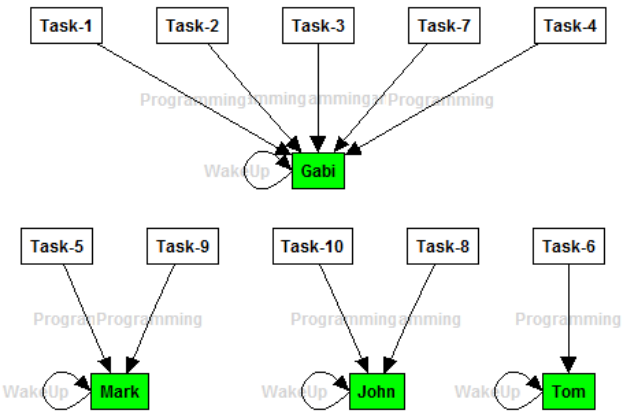


Fig. 9. The result of a particular SPSM simulation (assignment of programmers to tasks)

```
0,000   Anna   Ready   Programming     Busy    Task-1
1,782   Anna   Busy    WakeUp  Ready   Task-1
3,058   Anna   Ready   Programming     Busy    Task-2
4,380   Anna   Busy    WakeUp  Ready   Task-2
10,239  Anna   Ready   Programming     Busy    Task-11
10,332  Gabi   Ready   Programming     Busy    Task-12
10,598  John   Ready   Programming     Busy    Task-7
10,848  Mark   Ready   Programming     Busy    Task-5
12,431  Tom    Ready   Programming     Busy    Task-6
16,593  Mark   Busy    WakeUp  Ready   Task-5
16,749  Mark   Ready   Programming     Busy    Task-4
19,306  Gabi   Busy    WakeUp  Ready   Task-12
20,254  Gabi   Ready   Programming     Busy    Task-9
26,214  Mark   Busy    WakeUp  Ready   Task-4
30,358  Mark   Ready   Programming     Busy    Task-10
31,228  Gabi   Busy    WakeUp  Ready   Task-9
31,579  Tom    Busy    WakeUp  Ready   Task-6
34,636  Gabi   Ready   Programming     Busy    Task-3
36,639  Mark   Busy    WakeUp  Ready   Task-10
39,322  Mark   Ready   Programming     Busy    Task-8
41,086  Anna   Busy    WakeUp  Ready   Task-11
49,549  Gabi   Busy    WakeUp  Ready   Task-3
49,972  Mark   Busy    WakeUp  Ready   Task-8
58,962  John   Busy    WakeUp  Ready   Task-7
END
```

Fig. 10. The system's console with detailed information of task status, performer and dependencies

VIII. Examples of Ontorion applications

Due to the limitations of this paper, we only briefly mark a few Ontorion applications which we think give an objective perspective on its functionality.

First of all, our system has been successfully deployed as an intelligent semantic tool in a company from the energy sector located in the USA. There are several benefits worth mentioning of customizing Ontorion to this client. Primarily, we were able to semantically describe data sources due to automating the process of infrastructure management.

Another interesting application, which has taken place recently in a company in the Aeronautics and Space industry, is a case-based reasoning solution. Here, we combine text mining with a dedicated ontology to mine and structure information residing inside messages, incoming from users, which expose some third-party system errors and defects. Next, to those extracted chunks of information, the inference engine adds relevant tags based on a semantic analysis and together, such enhanced information (someone may even define such rich data as knowledge [33]) is exported to a database (knowledge base). Later, we execute triggers to combine this knowledge with rules and reason a possible set of actions. Nevertheless, an expert is responsible for selecting the final action to be taken due to the solution of the reported problem.

In medicine, for the Maria Skłodowska-Curie Institute of Oncology we built a highly complex ontology which represents a set of rules describing cancer procedure treatment. First and foremost, we are able to centrally manage all the rules, where, on a user's rule change request, with little effort, we just need to modify the semantic description preserving other rules from unnecessary modifications. The most beneficial is the usage of (semi) natural language that is readable for medical oncology experts. Thus in this way they are able to verify the knowledge input.

In the production sector, we have developed a common dictionary for different actors to communicate and collaborate world-wide. Actors include employees (located in different countries and continents) and heterogeneous information systems (from different software vendors). The deployment of semi-natural language, integrated inside the Fluent Editor (Ontorion ontology editor), proved to be a tool, on the one hand, easy to understand and use by its users, and on the other hand, easy to configure and maintain for administrators during systems (applications) integration.

Finally, we can use Ontorion to manage authorization and authentication, versioning, auditing and what distinguishes us from the competition is collaborative ontology engineering. Moreover, we are able to deploy a solution for semantic enhance searching which, based on taxonomy, efficiently improves sharing and searching for information in response to a user query, and interpreting strings of words not only using statistical techniques, but in the sense of logical connections existing between them. On the other hand, such a taxonomy can be seen as an asset of organization knowledge, which can be used to acquire knowledge from individuals and next to preserve it in a formal way.

IX. Future work directions

First, we plan to explore the scalability of MASS created in the way described in this paper. Even if Ontorion itself is a scalable solution, it is not obvious that the knowledge and inter-agent communication via a distributed queue will have the property of scalability too. Secondly, we want to explore an automated agent synthesis, based on theorem proofers. The synthesis should select rules in an adaptive way from the set of available rules by activating them with some threshold.

X. Conclusions

In this paper, we showed that the Ontorion server is able to execute, maintain and control massive simulations based on a hybrid MASS approach. The resulting MASS fits well into the definition of a Hybrid MASS. The perception of an agent is realized by description logic theorem proofers (reasoners). Agents are modelled as instances equipped with a relevant time model, which allows them to interact with the environment and with each other. Inter-agent communication is realized by messages that together with the environment are represented by an ontology managed by the server. Finally, actions performed by agents are encoded in a particular programming language, which brings our approach close to AOP.

An agent synthesis benefits from a formal reasoning engine (being a central component of KRR) and is based on an action-selection procedure. An expressive and distinct Ontorion functionality is the ability to encode agent logics in semi-natural language in the interest of a less professional user. We also observed that this allows end users to understand actions taken

by an agent, even if a user is not well trained in formal representation systems.

Obviously, we are aware of some obstacles which can be pointed out in the presented simulation. The key issue is to verify the usability and then estimate the degree of functionality adoption on the client side. We have also not used any technique of knowledge verification and validation [34]. On the other hand, we presented some Ontorion applications in the medicine, aerospace and production industries which were positively evaluated and are still in use. Based upon preliminary feedback from our clients, we think that the presented system, as a multi-agent simulation platform, is a promising prospect not limited to any particular industry or purpose.

Ontorion is free of charge for academic institutions and independent researches. For more information, please visit our website available at http://www.conitum.eu/semantics/.

REFERENCES

[1] D. Garvin, *Learning in Action: A Guide to Putting the Learning Organization to Work*. Harvard Business School Press, 2000. ISBN 9781578512515. [Online]. Available: http://books.google.co.nz/books?id=HAaZbow_cpUC

[2] J. A. Jakubczyc and M. L. Owoc, "Contextual knowledge granularity," in *Proceedings of Informing Science & IT Education Conference (InSITE)*, 2011, pp. 259–268.

[3] M. Bac, J. Korczak, A. Fafula, and K. Drelczuk, "A-trader - consulting agent platform for stock exchange gamblers." in *FedCSIS*, 2012, pp. 963–968.

[4] J. Korczak, M. Hernes, and M. Bac, "Risk avoiding strategy in multi-agent trading system." in *FedCSIS*, 2013, pp. 1119–1126.

[5] P. Weichbroth and M. Owoc, "A framework for web usage mining based on multi-agent and expert system an application to web server log files," *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, no. 206, pp. 139–151, 2011.

[6] C. Orłowski, A. Ziółkowski, and A. Czarnecki, "Validation of an agent and ontology-based information technology assessment system," *Cybernetics and Systems: An International Journal*, vol. 41, no. 1, pp. 62–74, 2010.

[7] A. S. Rao and M. P. George, "BDI agents: From theory to practice," in *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, 1995, pp. 312–319. [Online]. Available: http://www.agent.ai/doc/upload/200302/rao95.pdf

[8] S. Kripke, *Naming and necessity*. Harvard University Press, 1980. ISBN 9780674598461. [Online]. Available: http://bks0.books.google.com.ec/books?id=9vvAlOBfq0kC

[9] R. A. Brooks, "Intelligence without Reason," in *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, 1991, pp. 569–595.

[10] F. Bellifemine, A. Poggi, and G. Rimassa, "JADE - A FIPA-compliant agent framework," CSELT, Tech. Rep., 1999.

[11] I. A. Ferguson, "Touring machines: Autonomous agents with attitudes," *Computer*, vol. 25, no. 5, pp. 51–55, May 1992. doi: 10.1109/2.144395. [Online]. Available: http://dx.doi.org/10.1109/2.144395

[12] M. Woolridge and M. J. Wooldridge, *Introduction to Multiagent Systems*. New York, NY, USA: John Wiley & Sons, Inc., 2001. ISBN 047149691X

[13] D. Schmidt, M. Stal, H. Rohnert, and F. Buschman, *Pattern-Oriented Software Architecture: Patterns for Concurrent and Networked Objects*, ser. Wiley Series in Software Design Patterns. John Wiley & Sons, 2000, vol. 2.

[14] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, January 2003. ISBN 0521781760

[15] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 Web Ontology Language Primer," World Wide Web Consortium, W3C Recommendation, October 2009. [Online]. Available: http://www.w3.org/TR/owl2-primer/

[16] Cognitum. Ontorion Semantic Knowledge Management Framework. http://www.cognitum.eu/semantics/ontorion/. Made available on 20 March 2015.

[17] B. Quilitz and U. Leser, "Querying distributed rdf data sources with SPARQL," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds. Springer Berlin Heidelberg, 2008, vol. 5021, pp. 524–538. ISBN 978-3-540-68233-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68234-9_39

[18] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/sigops/sigops44.html#LakshmanM10

[19] P. Kaplanski, "Syntactic modular decomposition of large ontologies with relational database," in *ICCCI (SCI Volume)*, 2009, pp. 65–72.

[20] I. Horrocks, O. Kutz, and U. Sattler, "The even more irresistible sroiq." in *KR*, P. Doherty, J. Mylopoulos, and C. A. Welty, Eds. AAAI Press, 2006. ISBN 978-1-57735-271-6 pp. 57–67.

[21] G. Meditskos and N. Bassiliades, "Dlejena: A practical forward-chaining owl 2 rl reasoner combining jena and pellet," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, no. 1, 2010. [Online]. Available: http://www.websemanticsjournal.org/index.php/ps/article/view/176

[22] F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A distributed storage system for structured data," *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06)*, 2006.

[23] T. White, *Hadoop: The Definitive Guide*, first edition ed., M. Loukides, Ed. O'Reilly, june 2009. [Online]. Available: http://oreilly.com/catalog/9780596521981

[24] L. Lamport, "Paxos made simple, fast, and byzantine," in *OPODIS*, 2002, pp. 7–9.

[25] Cognitum, "Fluent Editor 2014 - Ontology Editor," http://www.cognitum.eu/semantics/FluentEditor/, made available on 20 March 2015.

[26] N. E. Fuchs, U. Schwertel, and R. Schwitter, "Attempto controlled english - not just another logic specification language," in *LOPSTR '98: Proceedings of the 8th International Workshop on Logic Programming Synthesis and Transformation*. London, UK: Springer-Verlag, 1990. ISBN 3-540-65765-7 pp. 1–20.

[27] M. Proctor, M. Neale, B. McWhirter, K. Verlaenen, E. Tirelli, A. Bagerman, M. Frandsen, F. Meyer, G. D. Smet, T. Rikkola, S. Williams, and B. Truit, "Drools," 2007. [Online]. Available: http://labs.jboss.com/drools/

[28] OMG. (2008) Semantics of business vocabulary and business rules (sbvr), v1.0. http://www.omg.org/spec/SBVR/1.0/PDF, Abruf am 02.05.2013. [Online]. Available: http://www.omg.org/spec/SBVR/1.0/PDF

[29] B. Glimm, M. Horridge, B. Parsia, and P. F. Patel-Schneider, "A syntax for rules in OWL 2." in *OWLED*, ser. CEUR Workshop Proceedings, R. Hoekstra and P. F. Patel-Schneider, Eds., vol. 529. CEUR-WS.org, 2008. [Online]. Available: http://dblp.uni-trier.de/db/conf/semweb/owled2009.html#GlimmHPP08

[30] M. I. Kellner, R. J. Madachy, and D. M. Raffo, "Software process simulation modeling: Why? what," *Journal of Systems and Software*, vol. 46, pp. 91–105, 1999.

[31] A. Czarnecki, C. Orłowski, T. Sitek, and A. Ziółkowski, "Information technology assessment using a functional prototype of the agent based system," *Foundations of Control and Management Sciences*, vol. 9, pp. 7–28, 2008.

[32] A. Czarnecki and C. Orłowski, "Ontology as a tool for the it management standards support," in *Agent and Multi-Agent Systems: Technologies and Applications*, ser. Lecture Notes in Computer Science, P. Jędrzejowicz, N. Nguyen, R. Howlet, and L. Jain, Eds. Springer Berlin Heidelberg, 2010, vol. 6071, pp. 330–339. ISBN 978-3-642-13540-8. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13541-5_34

[33] J. Gołuchowski, "Technologie informatyczne w zarządzaniu wiedzą w organizacji." *Prace Naukowe. Akademia Ekonomiczna w Katowicach*, 2007.

[34] M. A. Mach and M. L. Owoc, "Validation as the integral part of a knowledge management process," in *Proceeding of Informing Science Conference*, 2001.

13th Conference on Advanced Information Technologies for Management

WE ARE pleased to invite you to participate in the 11th edition of Conference on "Advanced Information Technologies for Management AITM'15". The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

Topics

The topics of interest include but are not limited to:

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

Event Chairs

Dudycz, Helena, Wrocław University of Economics, Poland

Dyczkowski, Mirosław, Wrocław University of Economics, Poland

Korczak, Jerzy, Wrocław University of Economics, Poland

Program Committee

Abramowicz, Witold, Poznan University of Economics, Poland

Ahlemann, Frederik, University of Duisburg-Essen, Germany

Andres, Frederic, National Institute of Informatics, Tokyo, Japan

Brown, Kenneth, Communigram SA, France

Chmielarz, Witold, University of Warsaw, Poland

Cortesi, Agostino, Università Ca' Foscari, Venezia, Italy

Czarnacka-Chrobot, Beata, Warsaw School of Economics, Poland

De, Suparna, University of Surrey, Guildford, United Kingdom

Dufourd, Jean-François, University of Strasbourg, France

Franczyk, Bogdan, University of Leipzig, Germany

Januszewski, Arkadiusz, UTP University of Science and Technology in Bydgoszcz, Poland

Kannan, Rajkumar, Bishop Heber College (Autonomous), Tiruchirappalli, India

Kersten, Grzegorz, Concordia University, Montreal, Poland

Kowalczyk, Ryszard, Swinburne University of Technology, Melbourne, Victoria, Australia

Kozak, Karol, Fraunhofer and Uniklinikum Dresden

Leyh, Christian, Technische Universität Dresden, Chair of Information Systems, esp. IS in Manufacturing and Commerce, Germany

Ligęza, Antoni, AGH University of Science and Technology, Poland

Ludwig, André, University of Leipzig, Germany

Magoni, Damien, University of Bordeaux – LaBRI, France

Michalak, Krzysztof, Wroclaw University of Economics, Poland

Owoc, Mieczysław, Wroclaw University of Economics, Poland

Pankowska, Malgorzata, University of Economics in Katowice, Poland

Pawełoszek, Ilona, Częstochowa Univeristy of Technology

Quirin, Arnaud, University of Vigo

Rot, Artur, Wroclaw University of Economics, Poland

Rudek, Radosław, Wrocław University of Economics

Stanek, Stanisław, General Tadeusz Kosciuszko Military Academy of Land Forces in Wroclaw, Poland

Surma, Jerzy, Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States

Teufel, Stephanie, University of Fribourg, Switzerland

Tsang, Edward, University of Essex, United Kingdom

Wolski, Waldemar, Uniwersytet Szczeciński

Zanni-Merk, Cecilia, Universite de Strasbourg, France

Ziemba, Ewa, University of Economics in Katowice, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 187–193
DOI: 10.15439/2015F320
ACSIS, Vol. 6

Knowledge representation in controlling sub-system

Anna Chojnacka-Komorowska
Wrocław University of Economics ul. Komandorska 118/120, 53-345 Wrocław, Poland
Email: anna.chojnacka@ue.wroc.pl

Marcin Hernes
Wrocław University of Economics ul. Komandorska 118/120, 53-345 Wrocław, Poland
Email: marcin.hernes@ue.wroc.pl

***Abstract*—The paper presents issues related to knowledge representation in controlling sub-system in integrated management support system. In the first part of article, a controlling sub-system is characterized. Next, the formal definition of knowledge structure is presented. This structure can be used, for example, to integration of knowledge. The final part describes an example of use of the elaborated structure in practice.**

I. Introduction

CONTEMPORARY economy forces companies to operate in a very turbulent environment. In order for the companies to become competitive, the decision makers are forced to make quick yet effective decisions. Without a doubt, the effectiveness of those decisions influences performance and results obtained by companies. To run the decision-making process correctly, companies increasingly employ the process of controlling, defined as performance-oriented operation, executed through planning, control and reporting [9]. Professional literature considers controlling in two aspects: strategic controlling which is aimed at coordinating all sub-systems of strategic management, i.e. strategic planning, control as well as strategic information feed [5]; and operational controlling which should put the company in control over its income, expenses, thus achieving the assumed profit, and financial liquidity [2]. In order to ensure full integration of business processes within a company, supporting the tasks of both strategic and operational controlling should be executed within a sub-system of an integrated management information system (IMIS) [1]. It should be noted that the function of controlling sub-system for analysis, planning or control generates new knowledge which should be used by the decision makers on an ongoing basis.

However, so far professional literature has not defined formal representation structure of the knowledge in a controlling sub-system, considering all elements of the controlling system. Such structure might prove useful, e.g. to compare items of knowledge generated with different methods of analysis.

Therefore, the purpose of this article is to devise a formal definition of knowledge structure in IMIS controlling sub-system.

The research has been realized through the following stages:

1. Analyzing of the existing solutions in knowledge representation and integration and controlling area using such research methods like the literature studies, the observation of phenomena in the enterprises, the case studies of different practical application of IMIS.
2. Developing the definition of the knowledge structure. The quantitative methods and the case studies have been used in this stage.

II. Knowledge representation and integration techniques

A knowledge representation structure must be capable of representing the broad spectrum of knowledge types categorized by Feigenbaum including [11]:

- objects - information on physical objects and concepts,
- events - time-dependent actions and events that may indicate cause and effect relationships,
- performance - procedure or process of performing tasks,
- meta-knowledge - knowledge about knowledge including its reliability, importance, performance evaluation of cognitive processors.

The literature of subject presents many different methods for knowledge representation. The main of them include internal-symbolic representation, first-order predicate logic, multi-valued logic, fuzzy logic, tuples, relational tuples, partitions and coverings, trees, rule-based systems, artificial neural networks, frame representation, ontologies such as semantic web and semantic networks, multi-attributes and multi-values structures, multi valued logic includes a three valued logic and a fuzzy logic.

Internal-symbolic representation requires a common symbol language, in which knowledge can be express.

The first-order predicate logic (and its extensions), multi-valued logic and fuzzy logic approach is a symbolic-

cognitive approach and results from general assumptions [14]:

- the knowledge representation is independent of physical media,
- system's internal states are related to the objects of external environment,
- the knowledge representation consist of symbols forming the structure,
- reasoning is based on the manipulation of these structures to derive other structures.

Often the tuples, relational tuples, partitions and coverings are used for knowledge representation [15]. Trees, instead, are the graphs which represent hierarchical knowledge [12], [26].

A rule based system uses rules as the knowledge representation for knowledge coded into the system. A rule-based system consists of a set of IF-THEN rules, a set of facts and some interpreter controlling the application of the rules, given the facts [13].

Artificial neural networks are generally defined as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature [18].

Frame representation provides a concise structural representation of useful relations, and support a concise definition-by-specialization technique that is easy for most domain experts to use. In addition, special purpose deduction algorithms have been developed that exploit the structural characteristics of frames to rapidly perform a set of inferences commonly needed in knowledge-system applications. the taxonomic relationships among frames enable descriptive information to be shared among multiple frames (via inheritance) and because the internal structure of frames enables semantic integrity constraints to be automatically maintained. [19].

The Semantic Web allows searching not only information but also knowledge. Its main purpose is introducing structure and semantic content in the huge amount of unstructured or semi-structured distributed knowledge available on the Web, being the central notion behind the Semantic Web that of ontologies, which describe concepts and their relations in a particular field of knowledge [16].

The paper [17] instead, presents the semantic net with node and links activation level (the "slipnet") to represent knowledge. This type of representation allows the processing both knowledge represented in a symbolic way and knowledge represented in a numerical way. Thus it is possible to determine a certainty level of semantic relations between nodes (topics).

Often knowledge is represented as multi-attribute and multi-value structure consist of different number of different types attributes. It allow for representing the real word environment in wide scope of objects features. Such structures are used, for example in case of weather forecasting multiagent system [15] or supply chain management multiagent system [10], [25].

Integration of knowledge is considered in different ways and on different levels. Previous attempts tended to predict group performance based on some statistic involving members' performances. For example [20] reported that group performance is an average of individual performance. Wolley [21] measured the collective knowledge. They experimentally proved that collective knowledge and intelligence is not strongly correlated with the average or maximum individual intelligences of group members but is correlated with the average social sensitivity of group members, the equality in distribution of conversational turn-taking, and the proportion of females in the group. In works [22], [23], [24] a formal mathematical model for knowledge integration is presented, in which the consensus-based knowledge functions for generating integration of knowledge have been defined.

III. CHARACTERISTIC OF A CONTROLLING SUB-SYSTEM

Operational management of a company requires employing a reliable and effective controlling system, which depends on proper organization of an integrated management system. Controlling sub-system is a component of IMIS, collecting big amounts of data from other sub-systems, thus generating information for the management. The information is then the basis for making a decision. Therefore, the reliability of the sub-system is dependent upon the reliability of the entire IMIS. It is also imperative to provide support for the controlling system through maximally automated circulation of documents, which allows for an ongoing data feed for management purposes; through a register of events taking place in a company, that enables to control, and through issuing up-to-date and solid analyses and reports for the management. Such solutions are made possible through the use of domain bases of each of the employed IMIS sub-systems, and the use of data warehouses along with other Business Intelligence tools. Summary of sample information used by the controlling sub-system in order to the planning function, control and inspection presented table 1.

Considering the controlling sub-system and its efficient operation, it is vital to decentralize the management system in a company and one of the ways to do that is to designate certain responsibility centers in the company [8]. The starting point for such division is an assumption that the entire activity conducted by the company can be divided into many integrated sub-categories, each of them having its own individual characteristic and requiring an individual approach [6]. Dividing a company into smaller units and assigning specific targets to them is a crucial condition for introducing a reliable controlling system and making unit managers account for the accomplishment of assigned tasks and targets [8].

Fig 1 presents the functional architecture of controlling subsystem. From the knowledge processing point of view,

TABLE I.
EXAMPLES OF THE USE OF EVENT RECORDS IN OTHER SUB-SYSTEM IN ORDER TO THE CONTROLLING FUNCTIONING IN ENTERPRISE

The name of the sub-system	Automatic records of events supporting the work of the controlling
ACCOUNTING SUB-SYSTEM	– automatic generation posting and transfer automatically after payment, – automatic posting costs and financial income resulting in exchange rate differences, – automatic decreeing and posting payroll generated in human resources management sub-system, – automatic aggregation and post inventory documents.
HUMAN RESOURCES MANAGEMENT SUB-SYSTEM	– information about the amount of the salaries paid in individual organizational units, – information about the differences in salaries on individual positions, – information about the differences in wages on individual positions, – information about skilled workers needed to perform specific jobs by company, – automatic creation of the collator salary costs on projects/organizations units on the basis of the records of working time.
LOGISTIC SUB-SYSTEM	– information on current inventory values, broken down by types of materials or their economic usefulness, – receiving/issuing of goods from the warehouse on the basis of the barcode (automatic records), – carry out inventory using barcodes or technology (Optical Character Recognition/ Intelligent Character Recognition/ Optical Mark Recognition), – automatic generation of an inventory documents.
CRM SUB-SYSTEM	– generate information about the current state of the debt, thanks to the automatic generation of invoices, – the ability to create reports based on customer segmentation or Pareto principle, – information on complaints made by customers.
MANUFACTURING MANAGEMENT SUB-SYSTEM	– identify needs of material of a specific production order, – determine necessary for the order execution the man-hours, – analysis of the use of the working time of machines and equipment on the basis of the automatic reading of their work time.

the reports module and internal consultant module are very important. The first of them facilitate monitoring of the business organization processes. It is very important tool for management supporting. After determining the limit values for a given process the module reports values of each event and informs about tasks, which have to be perform. The second module, instead, suggests the optimal solutions for a given decision-making problem. For example, the tasks performed by module are as follows:

- performing the effectivity analyses,
- supporting in the costs calculation rules,
- supporting in the costs settlement.

When implementing the controlling sub-system in a company, it is vital to remember the specific needs of each organization and their uniqueness. Therefore, it is necessary to precisely define targets and expectations put forth to the controlling process by management staff from each company at pre-implementation stage. This will allow obtaining an efficient tool, adapted perfectly to the information needs of a company. One can honestly state that there are no identical controlling systems, as there are no identical companies. Distinct characteristics of production processes, distinct organization methods and different management methods make each company one of a kind [3, 4].

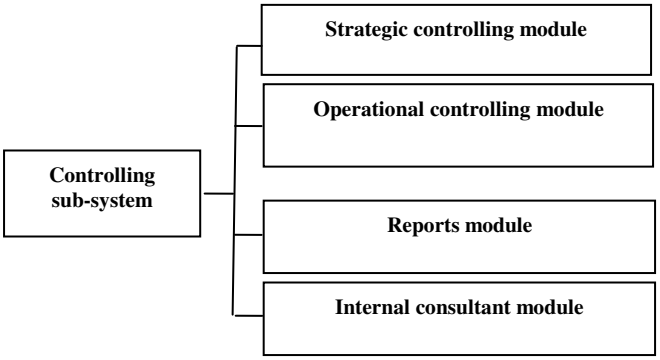


Fig 1. The functional architecture of a controlling sub-system.
Source: self-elaboration.

Consequently, each company has its own uniquely organized controlling system, although the system contains commonly used solutions and tools, based on two basic assumptions:

- all action taken in the company is aimed at the main target – to optimize profit made in a given environment,
- to increase the efficiency of all centers of responsibility, so that they contribute to achieving the main target of the company to the highest degree possible.

As a result of the analysis and conclusion process, the controlling sub-system generates new knowledge concerning business processes. The knowledge is subject to structuration process, therefore it should be represented as unitary structure which will be defined later in the article. The computer system should provide the controlling unit with all necessary information from sub-systems operating in the company, including the financial-accounting or production management sub-systems and allow analyzing received data in various lay-outs and cross-sections.

IV. DEFINITION OF THE KNOWLEDGE STRUCTURE

In order to allow for the representation of knowledge in controlling sub-system as a single structure, it is necessary to define it formally. On the basis of the characteristics of the sub-system control made in the previous paragraph, the knowledge structure can be defined as follows:

The structure of knowledge in controlling sub-system is called following sequence:

$$WCO = \langle \{D\}, \{P\}, \{W\}, \{AN\}, \{J\}, \{K\}, \{R\}, \kappa, \omega, SP, DT \rangle$$

Where:

1) $D = \{d_1, d_2, \ldots, d_i\}$ - denotes set of data, for example, record of production orders, a list of the materials ordered from vendors, list of cost items and revenue which undertaking the subject to the budgeting by each of the organizational units, or bill of the customers or vendors,

2) $P = \{p_1, p_2, \ldots, p_i\}$ - denotes set of plans prepared for a specific period, in scope of data of D set,

3) $W = \{w_1, w_2, \ldots, w_i\}$ – denotes set of values related to the implementation of the established size plans by individual organizational units operating in the enterprise,

4) $AN = \{a_1, a_2, \ldots, a_i\}$ - denotes set of analysis made on the basis of plans and implementation, and calculated on the basis of the analysis the values of deviations are presented in reports,

5) $K = \{k_1, k_2, \ldots, k_i\}$ - denotes set of controls, where $k_1 .. k_i$ mean procedures of controls, allow to detect irregularities in the use of production factors by, for example, comparison of standards with real values; an example would be control of fuel consumption in transport vehicles in relation to established standards of consumption, or quality control of finished products on each stage of production;

6) $J = \{j_1, j_2, \ldots, j_h\}$ – denotes set of organizational units to evaluate using the tools of controlling, with its manager about specific responsibilities and competence necessary to drive organizational unit,

7) $R = \{r_1, r_2, \ldots, r_h\}$ - denotes set of reports generated by system, both of the checks carried out and in terms of presentation of information resulting from the analysis carried out,

8) SP - denote s the degree of certainty of the reports, in particular, in relation to reports on the nature of the *„ex ante”*. Degree of certainty can be, for example, calculated on the basis of the probability of a change in interest rates by the central bank, changes in inflation rates and different sizes for the purchasing power of money,

9) $DT = \{CT, CW\}$ - where CT denotes transaction time, instead CW the right time of the reports performing, e.g. report performed 2013-01-23 showing an enterprise state of the day 2012-12-31.

In addition to the defined attributes of the knowledge structure includes the following features:

10) $\kappa : D \times P \times W \times K \to R$ - is at least partially a function of control, that mirrors elements of the Cartesian product $D \times P \times W \times K$ in elements of R set. Function κ will be partially, when only selected elements of the Cartesian product $D \times P \times W \times K$ will be as its arguments. This function, on the basis of the data, plans, implementation and control, creates a report with the controls.

Function κ satisfies the following conditions:

1) $(D = \varnothing) \vee (P = \varnothing) \vee (W = \varnothing) \vee (K = \varnothing) \Rightarrow \kappa = \varnothing$
   - if any of function's arguments is an empty set, then the function result is also an empty set.
2) $(D \neq \varnothing) \wedge (P \neq \varnothing) \wedge (W \neq \varnothing) \wedge (K \neq \varnothing) \Rightarrow \kappa \neq \varnothing$.
   - if each of function's arguments isn't an empty set, then the function result also isn't an empty set.

11) $\omega : D \times P \times W \times AN \to R$ - is at least partially a function of knowledge, that mirrors elements of the Cartesian product $D \times P \times W \times AN$ in elements of R set. Function ω will be partially, when only selected elements of the Cartesian product $D \times P \times W \times AN$ will be as its arguments. This function, on the basis of the data, plans, implementations and analysis, creates a report.

Function ω satisfies the following conditions:

1) $(D = \varnothing) \vee (P = \varnothing) \vee (W = \varnothing) \vee (AN = \varnothing) \Rightarrow \kappa = \varnothing$
   - if any of function's arguments is an empty set, then the function result is also an empty set.
2) $(D \neq \varnothing) \wedge (P \neq \varnothing) \wedge (W \neq \varnothing) \wedge (AN \neq \varnothing) \Rightarrow \kappa \neq \varnothing$
   - if each of function's arguments isn't an empty set, then the function result also isn't an empty set.

The knowledge structure of controlling sub-system defined in this article is a multi-attributes and multi-values structure (it consist of different types of attributes). This structure can be using to representation of knowledge generated as a result of analysis of implementation established in the plans of the values.

Next part of the paper presents using a defined structure in practice implementation.

V. THE EXAMPLE OF USING THE FORMAL DEFINITION OF KNOWLEDGE STRUCTURE

The use of the knowledge structure is illustrated in the example companies of financial industry that take care of the free of debts the hospitals. The knowledge structure, automatically generated by controlling sub-system, can, in this case, present itself as follows:

1) D={list of hospitals operating on an interesting enterprise area, list of hospitals with which the enterprise has established cooperation, list of costs and revenues necessary for the proper preparation of the company's budget}.

2) P={prepared plans for the period of the financial year, broken down by 4 calendar months. The plans consist of:
- sales budget prepared on the basis of both the hospitals with which you are already working, as well as new units, the enterprise should acquire in the indicated period,
- cost budget broken down by individual business units, i.e. the marketing department, dales department, accounting department, controlling department and the others,
- expected results associated with the free of debts the specific hospital}.

3) W= { the value corresponding to a plan, resulting from the actual implementation of the previously established plans}.

4) AN={analyses consist of comparison plans and implementation the individual scheduled values included in set W}.

5) K={the control concerning of liquidity loss prevention in connection with the activities of large outflows of funds in the form of a credit for hospitals and the cyclical (for example, monthly or quarterly) a relatively small influence on behalf of the bank account}.

6) J={marketing department, dales department, accounting department, controlling department, Hospital 1, Hospital 2….Hospital n}.

Therefore, a set J, in this example, is a set of organizational units of the enterprise, but also the hospitals, z with which the enterprise cooperates to the clearance of the efficiency of the investments carried out,

7) R={reports on the implementation of budget plans and the results of the analysis of deviations and evaluation report on funding opportunities further hospitals based on the company's own or with the possibility to organize a bank loans intended for this purpose }.

8) SP =0.7.

9) DT ={01.09.2012, 31.12.2012};

The enterprise is considering financing the business development based on hospitals 1, 2, 3 and 4 with 5-year bank loan. The 700 thousand loans will ensure the financial liquidity in the investment period. The detailed calculations are presented in the table 2.

The parameters of automatic function of control and its results may present as follows:

κ (planned receipts and cash outflows) = {planned cash receipts, collected at the moment funds along with the loan had entered for the purpose of carrying out the activities and the planned financial resources allocated to free of debts the hospitals; result of a control: planned receipts and expenses are possible provided under condition to ensure the operational functioning of the company and to cover the cost of raising capital};

Instead, parameters and results (in the form of reports, which can contain both the information and the conclusions drawn in automatically[1]) of the function of knowledge generated by the sub-system may present as follows (the number of the tables was given for the order by the authors – it is not generated by the sub-system):

ω(D, P,W,AN)= {results are presented by Table 2 and Table 3}.

TABLE 2. PLANNED CASH FLOWS ASSOCIATED WITH THE LENDING AND REPAYMENT INSTALLMENTS BY HOSPITALS

The name of project	Year	Month	Payment term	Cash flows [thousands]	Balance [thousands]
Loan				**700**	**700**
Hospital 1	2012	9	2012-09-03	-150	550
Hospital 2	2012	9	2012-09-08	-380	170
Revenues	2012	9	2012-09-15	45	215
Operating costs	2012	9	2012-09-30	-20	195
Financial costs	2012	9	2012-09-30	-10	185
Revenues	2012	10	2012-10-30	55	240
Hospital 3	2012	10	2012-10-03	-120	120
Operating costs	2012	10	2012-10-31	-17	103
Financial costs	2012	10	2012-10-31	-14	89
Revenues	2012	11	2012-11-15	66	155
Operating costs	2012	11	2012-11-30	-20	135
Financial costs	2012	11	2012-11-30	-20	115
Revenues	2012	12	2012-12-15	66	181
Hospital 4	2012	12	2012-12-25	-140	41
Operating costs	2012	12	2012-12-31	-20	21
Financial costs	2012	12	2012-12-31	-21	0

Conclusions: Following the presented juxtaposition, one might initially state that a company is able to fund the planned projects through the planned amount of bank loan. Significant financial expenses of the bank loan make the company management worried about the profitability of such investment. The revenues, expenses and profits in the described period are presented in the table 3.

[1] The data mining techniques and expert systems are used in this purpose

TABLE 3. PLANNED REVENUES, COSTS, PROFITS OF ENTERPRISE FUNCTIONING

	2012/09	2012/10	2012/11	2012/12
Revenues [thousands]	45	55	66	66
Costs [thousands]	30	31	40	41
Profit [thousands]	15	24	26	25
Cumulative Profit [thousands]	15	39	65	90

Conclusions: As one may notice the profit margin of the investment amounts to 38% and allows sustainable growth of the enterprise and repayment of the bank loan within agreed period.

Generating data from the chart by the controlling sub-system is possible due to its cooperation with the sales sub-system, because it contains all information regarding concluded agreements, their amount and payment schedule. The juxtaposition may function as final or be the basis for creating an overall plan of money flow through the controlling sub-system that considers all revenues, costs and profits of company's operation (table 3).

The similarly structures of knowledge can be automatically generated by the controlling sub-system with regard to, for example, other periods, or new hospitals acquired by the company. Thanks to the formal representation of knowledge it can be easily compared, verified and integrated.

VI. CONCLUSION

The operation of the controlling sub-system in IMIS is connected with generating knowledge which is extremely useful e.g. from the perspective of company competitiveness. Operational efficiency is not only influenced by flexibility and the ability to satisfy customers' needs, but also by keeping proper financial liquidity in the company. Note that economic decisions are usually made with risk and uncertainty, therefore knowledge generated by the controlling sub-system is often heterogeneous by nature. That is why it is crucial to store the knowledge with the use of unitary structure whose formal definition has been devised in this article. Representing knowledge with the use of such structure enables its comparison as well as detection of conflicts resulting from the knowledge – for instance, one can obtain a report from the controlling sub-system, which states that within a given period of time it will be more profitable for the company to produce product A for customer K1, however, production of product B for customer K2 within the same period of time will ensure higher financial liquidity. Such a situation is defined as conflict of knowledge which hinders the process of making a final decision, and should be resolved by man or by an automated system [25].

The representation of knowledge in the controlling sub-system as unitary structure may thus lead to higher effectiveness of all decisions made by decision makers, based on reports generated in the controlling sub-system.

The further research works may relate to elaborate a structure of knowledge with functional dependencies between attributes, formal definition of knowledge conflict appearing in controlling sub-system and the methods of these conflicts resolving.

REFERENCES

[1] A. Bytniewski (ed.), "Architektura zintegrowanego systemu informatycznego zarządzania", Wydawnictwo AE we Wrocławiu, Wrocław 2005.

[2] A. Chojnacka – Komorowska, "Projektowanie rozwiązań controllingu operacyjnego w przedsiębiorstwie", in: Bytniewski A. (ed.) *Informatyka ekonomiczna. Informatyka w biznesie.* Prace Naukowe UE nr 159, UE, Wrocław 2011.

[3] I. Chomiak-Orsa, "Wykorzystanie nowoczesnych technologii w doskonaleniu procesów controllingowych", *Informatyka Ekonomiczna,* Prace Naukowe AE nr 1150, Wrocław 2007.

[4] E. Baraldi, M. Ingemansson, and A. Launberg, "Controlling the commercialisation of science across inter-organisational borders: Four cases from two major Swedish universities", *Industrial Marketing Management,* 43(3) 2014, pp. 382-391, doi:10.1016/j.indmarman.2013.12.006

[5] A. Januszewski, "Systemy systemy rachunkowości i controllingu", in: Zawiła-Niedźwiecki J., Rostek K., Gąsiorkiewicz A. (ed.), *Informatyka Gospodarcza*, tom 2, Wydawnictwo C.H. Beck, Warszawa 2010.

[6] J.Y. Wang, "Controlling shareholder entrenchment: Bonuses versus dividends", *International Review of Economics and Finance* 32, 2014, pp. 143-158. doi:10.1016/j.iref.2014.01.012

[7] S. Marciniak, "Controlling. Teoria zastosowania", Diffin, Warszawa 2008.

[8] J. Nesterak, "System oceny centrów odpowiedzialności", ANVIX, Kraków 2004.

[9] S. Nowosielski, "Zarządzanie produkcją – ujęcie controllingowe", Wydawnictwo AE we Wrocławiu, Wrocław 2001.

[10] J. Sobieska-Karpińska, and M. Hernes, "Consensus determining algorithm in multiagent decision support system with taking into consideration improving agent's knowledge", *Proceedings of the Federated Conference on Computer Science and Information Systems* (FedCSIS), 2012.

[11] P. Tanwar, T. V Prasad, and K. Datta , "Hybrid technique for effective knowledge representation & a comparative study" CoRR abs/1209.3869, 2012. doi: 10.1007/978-3-642-31600-5_4

[12] M. Maleszka and N.T. Nguyen, "Integration computing and collective intelligence", *Expert Systems with Applications*, vol. 42 (1), , 2015, pp. 358-378. doi:10.1016/j.eswa.2014.07.036

[13] C. Grosan, and A. Abraham, "Intelligent Systems - A Modern Approach", Springer-Verlag Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-21004-4

[14] J. Ferber, "Multi-Agent Systems", Addison-Wesley Longman 1999.

[15] N. T. Nguyen, "Using Consensus Methodology in Processing Inconsistency of Knowledge", [in] Last M. et al. (Eds): *Advances in Web Intelligence and Data Mining*, series Studies in Computational Intelligence, Springer-Verlag, 2006, pp. 161-170.

[16] Z. Zeng, "Construction of knowledge service system based on semantic web", *Journal of The China Society For Scientific and Technical Information*, 2005,24(3):336-340.

[17] S. Franklin and F.G. Patterson, "The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent". in: *Proc. of the Int. Conf. on Integrated Design and Process Technology*. San Diego, CA: Society for Design and Process Science, 2006.

[18] O. Badawy, and A. Almotwaly, "Combining neural network knowledge in a mobile collaborating multi-agent system", *Electrical, Electronic and Computer Engineering*, 2004. ICEEC '04. 2004 International Conference on , vol., no., pp.325,328, 5-7 Sept. 2004, doi: 10.1109/ICEEC.2004.1374457

[19] G.J. Zhu, and Y.M. Xia, "Research and practice of frame knowledge representation", *Journal of Yunnan University* (Natural Sciences Edition), 2006,28(S1):154-157.

[20] M.R. Barrick, G.L. Stewart, M. J. Neubert, and M.K. Mount, "Relating Member Ability and Personality to Work-Team Processes and Team Effectiveness", *Journal of Applied Psychology*, 83 (3), 1998, s. 377-391.

[21] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, "Evidence for a Collective Intelligence Factor in the Performance of Human Groups", *Science*, 330 (6004), 2010, s. 686-688.

[22] T.H. Duong, N.T. Nguyen, and G.S. Jo, "A Method for Integration of WordNet-based Ontologies Using Distance Measures", in: *Proceedings of KES 2008*. Lecture Notes in Artificial Intelligence 5177, 2008, pp. 210-219. doi: 10.1007/978-3-540-85563-7_31.

[23] L. Sliwko, and N. T. Nguyen, "Using Multi-agent Systems and Consensus Methods for Information Retrieval in Internet", *International Journal of Intelligent Information and Database Systems* 1(2), 2007, pp. 181-198. doi>10.1504/IJIIDS.2007.014949

[24] N.T. Nguyen, "Using consensus methods for solving conflicts of data in distributed systems", in: *Proceedings of SOFSEM 2000*, Lecture Notes in Computer Science 1963, 2000, pp. 411-419. doi: 10.1007/3-540-44411-4_30

[25] M. Hernes and J. Sobieska-Karpińska, "Application of the consensus method in a multiagent financial decision support system", *Information Systems and e-Business Management*, Springer Berlin Heidelberg 2015, doi: 10.1007/s10257-015-0280-9.

[26] M. Maleszka and N.T. Nguyen, "Approximate Algorithms for Solving O1 Consensus Problems Using Complex Tree Structure", *Transactions on Computational Collective Intelligence 8*, 2012, pp. 214-227. doi: 10.1007/978-3-642-34645-3_10

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 195–202
DOI: 10.15439/2015F319
ACSIS, Vol. 6

The semantic method for agents' knowledge representation in the Cognitive Integrated Management Information System

Marcin Hernes
Wrocław University of Economics
ul. Komandorska 118/120, 53-345
Wrocław, Poland
Email: marcin.hernes@ue.wroc.pl

***Abstract*—This paper presents a method for agents' knowledge representation by using semantic network with node and links activation level defined on the instance, concept, relation and axiom level. The first part shortly presents the state-of-the-art in the considered field; next, the CIMIS prototype is shortly characterized; the formal definition of a method for agents' knowledge representation is presented in the last part of paper.**

I. Introduction

CONTEMPORARY the entire economy is based on information and knowledge, therefore companies must employ systems which support the knowledge management process taking into consideration the risk and uncertainty of economic decisions. Often the integrated management information systems (IMIS) are used in this purpose. They are characterized by full integration both at the system/application level and the business process level. Note, however, that the properties of contemporary IMIS are becoming more and more inadequate. Apart from collecting and analyzing data and generating knowledge, the system should also be able to understand the meaning of phenomena occurring around the organization. It is becoming more and more necessary to make decisions based not only on knowledge but also on experience, thus far regarded as purely human domain [4]. In order to accomplish tasks set by IMIS, a multi-agent system can be used consist of several cognitive agents. Not only do they enable quick access to information and quick search for the required information, its analysis and conclusions, but also, besides being responsive to environment stimuli, they have cognitive abilities that allow them to learn from empiric experience gained through immediate interaction with their environments [15], which consequently allows a number of decision versions to be automatically generated and to make and execute decisions.

As a result of its running, the agent obtains knowledge of the environment in which it operates. If we desire to use this knowledge then it must be represented in the form of a specific structure. The ability of cognitive agents to understand the meaning of phenomena occurring around the organization causes that the semantic structures have to be used. The ontologies which are mainly applied for this purpose include semantic networks. They are formally defined in numerous papers (e.g. [9], [19], [27], [29]), however to a small extent they include the method for representation of the uncertainty of economic decisions. Therefore, the semantic net with nodes and links activation level is a better solution to represent the companies' knowledge. This type of representation enables processing both, knowledge represented in a symbolic way, and knowledge represented in a numerical way. It also enables processing structured and unstructured knowledge. Thus, it is possible to determine the certainty level of nodes (concepts) and semantic relations between these nodes. The first suggestion to use such a structure, called "slipnet", is presented in the "Copycat" project [14]. However, the existing papers lack the formal definition of this structure and do not consider issues related to instances and relations between nodes and instances. This definition is however necessary mainly in order developed methods for the agents' knowledge processing and integration.

The aim of this paper is to develop a formal definition of a method for agents' knowledge representation using semantic network with node and links activation level taking into consideration the instance, concept, relation and axiom (relations between nodes and instances) level. This structure has been implemented in the architecture of the cognitive agents running in the Cognitive Integrated Management Information System (CIMIS) prototype.

This paper is organized as follows: the first part shortly presents the state-of-the-art in the considered field; next, the CIMIS prototype is shortly characterized; the formal definition of a method for agents' knowledge representation is presented in the last part of paper.

II. RELATED WORKS

In this section the methods related mainly to cognitive agents' knowledge representation will be analyzed. In the study [6] considering the taxonomy of cognitive agent

This work was financially supported by the National Science Center (decision No. DEC-2013/11/D/HS4/04096)

architectures with respect to their knowledge representation and learning methods, three main groups of the architectures were distinguished:

1. Symbolic architectures which use declarative knowledge included in relations recorded at the symbolic level, focusing on the use of this knowledge to solve problems.
2. Emergent architectures using signal flows through the network of numerous, mutually interacting elements, in which emergent conditions occur, possible to be interpreted in a symbolic way.
3. Hybrid architectures which are the combinations of the symbolic and emergent approach, combined in various ways.

The literature of subject presents many different methods for agents' knowledge representation used in the mentioned groups. The main of them include first-order predicate logic, production systems, artificial neural networks, frame representation, ontologies such as semantic web, semantic networks and topic maps, multi-attributes and multi-values structures, multi valued logic includes a three valued logic and a fuzzy logic. Some of these methods are closely related to the semantic agents' knowledge representation.

Production systems are based on the production rules consisting of two elements: consequence or the head of the rule and the other is the antecedence or the body of the rule which should be true to satisfy the consequence part [16], [23].

Artificial neural networks are generally defined as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature [1], [18]. Actions taken by the agent are directly connected with perception without the mediation of symbolic reasoning. However, an area of application is limited to agents conceptually simple recognizing actions or underlying actions directly related to the recognition (for example, grasping objects).

It often happens that agents' knowledge is represented as multi-attribute and multi-value structure consisting of different number of different types of attributes. It allow for representing the real word environment in a wide scope of objects' features. Such structures are used, for example in case of the weather forecasting multiagent system [21] or supply chain management multiagent system [13], [25].

Frame representation captures the way agents typically think about Special Section of their knowledge, provide a concise structural representation of useful relations, and support a concise definition-by-specialization technique that is easy for most domain experts to use. In addition, special purpose deduction algorithms have been developed that exploit the structural characteristics of frames to rapidly perform a set of inferences commonly needed in knowledge-system applications. The taxonomic relationships among frames enable descriptive information to be shared among multiple frames (via inheritance) and because the internal structure of frames enables semantic integrity constraints to be automatically maintained. [8], [28].

Ontology, in turn, most often is defined by the following elements [9], [10]

$$O = \langle C, I, R, Z \rangle \tag{1}$$

where:

- C – Set of concepts (classes)
- I – Set of instances of concepts
- R – Set of binary relations defined on C
- Z – Set of axioms which are formulae of first-order logic and can be interpreted as integrity constraints or relationships between instances and concepts, and which cannot be expressed by the relations in set R, nor as relationships between relations included in R[1].

The Semantic Web allows searching not only information but also knowledge. Its main purpose is introducing structure and semantic content in the huge amount of unstructured or semi-structured distributed knowledge available on the Web, being the central notion behind the Semantic Web that of ontologies, which describe concepts and their relations in a particular field of knowledge [27].

Semantic networks [26] are knowledge representation schemes involving nodes and links (arcs or arrows) between nodes. The nodes represent objects or concepts and the links represent relations between nodes. The links are directed and labeled; thus, a semantic network is a directed graph. In print, the nodes are usually represented by circles or boxes and the links are drawn as arrows between the circles.

As an ontological knowledge representation, the topic map standard, introduced by International Organization for Standardization (ISO/IEC 13250:2000), is also used. The topic maps are a kind of a semantic network, and they allow writing information of the data ontology and data taxonomy in a semantically ordered manner [17]. The topic map, most often, consists of „parent-child" relations.

Presented methods for knowledge representation can be used only in respect of the one group of cognitive agent representation. For example, the topic maps can be used in symbolic architectures and neural networks can be used in emergent architectures. In the hybrid architectures often a combination of different methods is used depending on a task, which is to be executed. For example, if agents' purpose is visual quality analysis of grains, the neural network is used for recognizing the features of grains (e.g. shape, color) and semantic network is used to determine the grain species. However, the disadvantages of this approach are the need to implement two (or more) different types of modules for knowledge storing and the complexity of the procedures for conversion of knowledge represented using neural network with the knowledge represented by a

[1] This definition differs from that of "axioms" in generative grammar and formal logic. In ontology disciplines, axioms include only statements asserted as a priori knowledge.

semantic network. Therefore, a better approach is to use methods that allow representation of both symbolic and numerical knowledge in an integrated, uniform manner. The first suggestion to use such a method, called "slipnet", is presented in the "Copycat" project. This method is developed in the LIDA cognitive agent [7]. This hybrid architecture allows for symbolic and emergent knowledge processing and it uses the semantic net with node and links activation level (the "slipnet") to represent knowledge. This type of representation enables processing knowledge represented in a symbolic way, as well as knowledge represented in a numerical way. Thus, it is possible to determine a certainty level of semantic relations between nodes (topics).

Because the formal, mathematical definition of "slipnet" has not yet been defined and existing papers do not consider issues related to instances and relations between nodes and instances, it is very difficult to develop methods for advanced processing and integration of knowledge representing by this structure in CIMIS prototype, presented in the next part of this paper.

III. THE COGNITIVE INTEGRATED MANAGEMENT INFORMATION SYSTEM

In order to develop a prototype of CIMIS the following subsystems have been created (Fig 1.): fixed assets, logistics, manufacturing management, human resources management, financial and accounting, controlling, business intelligence [2].

The fixed assets sub-system includes support for the realization of processes related to fixed asset and involved their depreciation.

The logistics sub-system has all the main features supporting the employees of logistics department in their effective work [10]. The logistics sub-system enables maintaining optimal stock to meet the needs of production department.

The manufacturing management sub-system support a processes related to a manufacturing execution. It include functions from the scope of the technical preparation of production capacity, production planning, material consumption planning, planning and execution of a manufacturing tasks, manufacturing control, visualization, monitoring and archiving.

The human resources management sub-system supports realization of such processes, as the employees of the company data and contract registering, recording of working time, wage calculation, creating the tax and social security declaration.

The financial-accounting sub-system supports registering, to the full extent, economic events, also provides important, from the point of view of business management, information, concerning, inter alia, payment capacity, revenues, costs, financial result.

Controlling sub-system is automatically processing data related to profit and loss account in cooperation with accounting sub-system. The controlling sub-system consists of both a strategic and operational controlling.

The CRM sub-system is engaged in matters connected with ensuring the best company-customer relations and collecting information in the customers' preferences in terms of product purchase in order to increase sales. The enterprise's environment monitoring is also realized by this sub-system.

The purpose of business intelligence sub-system is to enable easy and safe access to information in a company, operation of its analysis and distribution of reports within the company and among its business partners, which in turn enables quick and flexible decision making. In the context of, most of all, the business intelligence sub-system, but other sub-systems as well, the CIMIS makes cognitive visualization [36] features available, meaning it enables a visualization of multi-dimensional data in one picture that allows finding the source of a problem in a short time and contributes to creating new knowledge about an object or problem [25].

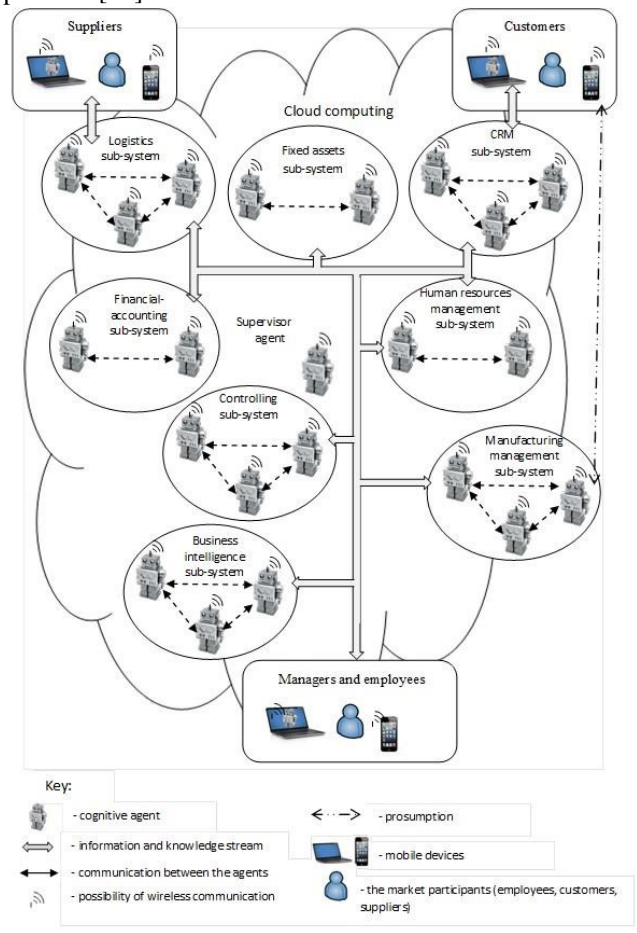


Fig 1. The CIMIS architecture.
Source: own work.

The system assumes that all agents are at 'not-taught' status in the initial phase. They can be initially grouped according company's needs for sub-systems. For in-stance, one group of agents is assigned to the logistics sub-system, another one is assigned to the manufacturing management sub-system and yet another one to financial and accounting sub-system. Within the groups, the agents can be initially 'taught' by the company that implements the system. Next stages of learning for both grouped and ungrouped agents are done by the company staff. Agents can also learn without teacher through analyzing the results of their decisions.

The agents of all sub-systems cooperate themselves in order to better business processes realization. For example, the enterprise's environment monitoring results performed by CRM sub-system agent are using by the other agents.

The main operating purpose of the Supervisor agent is to monitor the proper operation of other agents, mainly in the field of detection and solving conflicts of knowledge and experience. The agent analyzes, in close-to-real time, the structures of knowledge and experience of all agents. Whenever a conflict occurs, it employs a solution algorithm based on a method that uses consensus theory [9, 21], and the result of the agent's actions is accepted by the system as current state of knowledge and experience.

Note that all CIMIS sub-systems are connected by a single, coherent stream of in-formation and knowledge available online to the management, because nowadays attention is paid to functional complexity, managing all fields of operation in a company, proper flow of information and knowledge among sub-systems as well as the ability to perform a variety of analyses and to create reports for management. The implementation of this solution is realized as follow:

1. Communication between modules of agents architecture was ensured by using LIDA framework's codelets,
2. Communication between agents is based on Java Message Service (JMS) technology. The representation of information and knowledge (generated in result of agents' operating) in form of XML format document, was adopted (the JMS messaging is at the text type). The communication is realized in publish/subscribe messaging domains – it guarantees that information or knowledge generated by one of agents is immediately available for the other agents. The asynchronous message consumption is used.

All of the sub-systems functions are available as a local services or e-services (e.g. e-business, e-procurement, e-payment) by using Web Services technology.

At the physical level, the IMIS is built on the basis of the main two technologies – the LIDA framework (due to framework is developed at Java language and it is open the implementation of the other Java technologies – mentioned JMS, Java Database Connectivity or Java API for XML Web Services - is possible) and Microsoft SQL Server 2008 database management system.

The realization of the CIMIS is based on the LIDA cognitive agent architecture [7, 24], which is of emergent-symbolic nature, owing to which the processing of both structured and unstructured knowledge is possible.

In the LIDA architecture it was adopted that the majority of basic operations are performed by the so-called codelets, namely specialized, mobile programs processing in-formation in the model of global workspace. The functioning of the cognitive agent is performed within the framework of the cognitive cycle and it is divided into three phases: the understanding phase, the consciousness phase and the selection of actions and learning phase. At the beginning of the understanding phase the stimuli received from the environment activate the codelets of the low level features in the sensory memory [4]. The outlets of these codelets activate the perceptual memory, where high level feature codelets supply more abstract things such as objects, categories, actions or events. The perception results are transferred to workspace and on the basis of episodic and declarative memory local links are created and then, with the use of the occurrences of perceptual memory, a current situational model is generated; it other words the agent understands what phenomena are occurring in the environment of the organization. The consciousness phase starts with forming of the coalition of the most significant elements of the situational model, which then compete for attention so the place in the workspace, by using attentional codelets. The contents of the workspace module are then transferred to the global workspace (the so-called "broadcasting" takes place), simultaneously initializing the phase of action selection. At this phase possible action schemes are taken from procedural memory and sent to the action selection module, where there compete for the selection in a given cycle. The selected actions activate sensory-motor memory for the purpose of creating an appropriate algorithm of their performance, which is the final stage of the cognitive cycle [3]. The cognitive cycle is repeated with the frequency of 5–10 times per second.

Parallelly with the previous actions the agent's learning is performed, which is divided into perceptual learning concerning the recognition of new objects, categories, relations; episodic learning which means remembering specific events: what, where, when, occurring in the working memory and thus available in the awareness; procedural learning, namely learning new actions and action sequences needed for solving the problems set; conscious learning relates to learning new, conscious behaviors or strengthening the existing conscious behaviors, which occurs when a given element of the situational model is often in the workspace. The agent's learning may be performed as learning with or without a teacher.

It is worth emphasizing that each cognitive agent supporting decision-making must have the ability of

grounding the symbols, namely assign relevant real world objects to specific symbols of the natural language. This is necessary to correctly process unstructured knowledge saved mainly by means of the natural language and thus, for instance, the clients' opinions on products. The knowledge of this type is currently becoming more and more significant for a company because it may have impact on its competitiveness level. For instance analyzing the clients' opinions on a given product, the sales volume of a given product in the future may be estimated (of course with a certain level of probability).

The Cognitive Computing Research Group established by S. Franklin, developed in 2011 the framework (in Java language) significantly facilitating the implementation of the cognitive agent in CIMIS. It should also be emphasized that the whole framework code is open, i.e. the developer has access to the definitions of all methods. The learning mechanisms are not implemented in current version of the framework (they are under implementation by CCRG). This framework, however, does not contain the mechanism for automatically storing the agent's knowledgebase in a physical database. After the power is turned off, the agents' knowledgebase is lost. Initially the storing of an agents' knowledgebase in database has been launched after realizing by given codelet its task. However, this method proved to be insufficient in case more complex tasks. Therefore the need appears for developing a method for permanent, automatically storing the agent's knowledgebase in a physical database.

The next part of article presents a formal, mathematical definition of a semantic method for agents' knowledge representation based on semantic net with node and links activation level which includes instances and relations between nodes and instances.

IV. THE METHOD FOR AGENTS' KNOWLEDGE REPRESENTATION

The main ideas of using semantic network as agents' knowledge representation are as follows [5]:

- the meaning of a symbol or concept stems from relationships with other symbols and concepts; the human memory is a network of associations,
- information is contained in the nodes and arcs (links) connecting the nodes (node = concept; in the brain it is a pattern of beats activity of many neurons),
- every concept is a network node,
- the links between nodes are clearly presented,
- the links can be of different types,
- the semantic network is a model of episodic memory, but also semantic memory,
- the nodes represent, among other things: objects, types, or classes, events, activities, episodes, places, times,
- links represent, among other things: to give an example, subclass, the is-a relationship, it is part of something, logical conjunctions and, or, actions, instruments.

The need for agents' knowledge representation by semantic net with node and links activation level results mainly from the following presumptions:

- in a human brain from phonology and graphemes of the word to its meaning and model of the situation, we have different patterns of distribution (levels) of stimulation (activation), and associations between them [5],
- Pulvermuller [22] states, that because semantic activation followed by 90 ms the phonological activation then a brain stimulation is a natural base of semantic representation;
- very important issue is probability distribution (activation level); concepts related to the same topic better fit together and create a coherent concept graph of an active part of semantic memory including the inhibition of the activation and propagation.
- together with nodes and links and their activation level the instances and axioms have to be included in the semantic network; it facilitate an automatic storing an agents' knowledgebase in the physical database.

Taking into consideration: presented presumptions, the definition of ontology presented in section 2, and a "slipnet" features, the method for agents' knowledge representation, called "slipnetplus" is defined as follows:

Definition 1

The "slipnetplus" is called a quadruple:

$$SN = \langle N, I, L, Z \rangle \tag{2}$$

where:

N – set of nodes,

I – set of instances of nodes,

L – set of links i.e. set of fuzzy relations defined on N,

Z – set of axioms.♦

This definition extends the "slipnet" presented by [14, 24] about the set of instances and the set of axioms.

Let us to define the particular elements of "slipnetplus".

We assume a real world $<O, V>$ where O is a finite set of objects and V is the domain of O; that is, V is a set of object values, and

$$V = \bigcup_{o \in O} V_o \tag{3}$$

where V_o is the domain of object o.

We consider the "slipnetplus" referring to the real world (O, V) - such "slipnetplus" is called $<O, V>$-based. The "slipnetplus" detailed definitions must be considered on the four levels:

- the node level.
- the instance level.
- the link level.
- the axiom level.

These definitions are developed in the next subsections of this paper.

A. The node level

Definition 2.

A node of an <*O*, *V*>-based "slipnetplus" is defined as a triple:

$$Node = \langle n, O^n, V^n \rangle \tag{4}$$

where n is the unique name of the node, $O^c \in O \times [0,1]$ is a fuzzy set of objects represented by node with a certain level of probability, and $V^c \in V \times [0,1]$ is the objects' domain:

$$V^c = \bigcup_{\langle o,v \rangle \in O^n} V_o \times [0,1] \tag{5}$$ ♦

Nested pair $\langle O^n, V^n \rangle$ is called the structure of node *n*. It is obvious that all nodes belonging to the same "slipnetplus" are different from each other. However, notice that within a "slipnetplus" there may be two or more nodes with the same structure. Such a situation may take place, for example, for nodes "person" and "body". For expressing the relationship between them the links from set *L* will be very useful.

Set *N* in the "slipnetplus" definition is a set of nodes names and their activation levels.

B. The instance level

Definition 3

An instance of a node *n* is described by the objects from set O^n with values from set V^n and is defined as a pair:

$$instance = \langle i, v \rangle \tag{6}$$

where *i* is the unique identifier of the instance in world <*O*, *V*> and *v* is the value of the instance a tuple of type O^n, and can be presented as a function:

$$v : O^n \times [0,1] \to V^n \times [0,1] \tag{7}$$

such that $v(o, p) \in V_o$ for all $\langle o, p \rangle \in O^n$. ♦

Value *v* is also called a description of the instance within an object. A node may be interpreted as a set of all instances described by its structure.

We can then write $i \in n$ for presenting the fact that *i* is an instance of node *n*.

All instances of the same nodes within a "slipnetplus" should be different from each other. The same instance may belong to different nodes and may have different values.

C. The link level

In a "slipnetplus" within a pair of nodes there may be defined one or more links. Links between nodes describe the relationships between them. For example, between two nodes may be defined such relations as Synonym relation or Antonym relation. Links between nodes are included in set *L* of the "slipnetplus" definition.

Definition 4.

Let set *N* of nodes is given. The link is called the following relation:

$$L : N \times N \to [0,1] \tag{8}$$

in a space $N \times N$.♦

D.The axiom level

The set *Z* of axioms are formulae of fuzzy logic and can be interpreted as integrity constraints or relationships between instances and nodes, and which cannot be expressed by the relations in set *L*.

Definition 5.

Let set *N* of nodes and set *I* of instances are given. The axiom is called the following relation:

$$Z : N \times I \to [0,1] \tag{9}$$

in a space $N \times I$.♦

E. A graphical representation and an implementation of the "slipnetplus"

The developed definition of "slipnetplus" can be visualized in a graphical form. The Fig 2 presents an example of a graphical representation of the "slipnetplus".

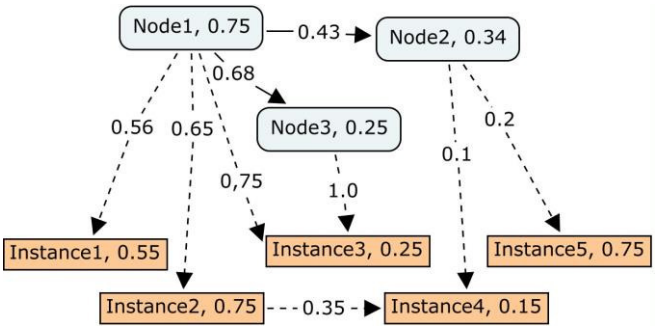


Fig 2. The example of a graphical representation of the "slipnetplus". Source: own work.

The arrows drawn with a continuous line represent links, while the arrows drawn dotted lines denote the axioms. The presented "slipnetplus" consist of three nodes with their activation levels. The interpretation of the *Node1* is as follows: The *Node1* exists in the real world with a probability level 0.75. Interpretation of other nodes is similar. The *Node1* is connected with *Node2* and *Node3* by links with the level of probability respectively 0.43 and 0.48.

The interpretation of the *Instance1* is as follows: The *Istance1* exists in the real world with a probability level 0.55. Interpretation of other instances is similar. The *Node1* is connected with *Instance1* by axiom with the level of probability 0.55. The *Node1* is connected with *Instance3* by axiom with the level of probability 0.75 and *Node3* is also connected with the Instance3 but with the level of probability 1.0.

The "slipnetplus" presented on Fig 2. can represent, for example, following practical economic situation:

Let:

- *Node1* denotes an "Investment",
- *Node2* denotes a "Securities",
- *Node2* denotes a "Currencies"
- *Instance1* denotes a "Gold",
- *Instance2* denotes a "Estate",

- *Instance3* denotes "EUR",
- *Instance4* denotes "Company1",
- *Instance5* denotes "Company2".

In considered situation it is necessary to invest on the 0.75 probability level[2]. The accessibility of "Securities" equals 0.34[3] and accessibility of "Currencies" equals 0.25. It is better to invest in "Currencies" (probability level 0.68) than in "Securities" (probability level 0.43). The accessibility of: "Gold" equals 0.55, Estate equals 0.75, "EUR" equals 0.25, "Company1" security equals 0.15 and "Company2" security equals 0.75. The axioms denote that the probability to achieve a positive (satisfactory) rate of return in case investment in the "Gold" equals 0.56 and in case securities of "Company2" equals 0.2. The interpretation of remaining axioms is similar.

Taking into consideration the implementation issues, the "slipnetplus" has been implemented in CIMIS. The code of LIDA framework classes related to "slipnet" implementation has been extended to the "slipnetplus" implementation. The LIDA agents' knowledgebase is automatically stored in a database by using instances. The object-oriented noSQL database model is suitable for storing nodes, links, instances, axioms together with their activation levels.

V. Conclusion

The formal, mathematical method for agents' knowledge representation uses semantic net with node and links activation level taking into consideration instances and relations between nodes and instances, has been developed in this paper. This method has been implemented in CIMIS and can be directly implemented in other multiagent systems. It should be noted, that this type of representation allows processing both knowledge represented in a symbolic way, and knowledge represented in a numerical way. Simultaneously a structured and unstructured knowledge can be processed. Thus, it is possible to determine a certainty level of semantic relations between nodes (topics). In case of the economic knowledge, it is a very important issue because the decisions-making process based on the type of knowledge usually takes place in conditions of risk and uncertainty.

The implementation of the developed semantic method for agents' knowledge representation greatly facilitated the agents' knowledge base mapping in the physical database.

The developed method enables the realization of further research works related to developing methods for processing and integration knowledge represented by the "slipnetplus". Also action selection and action performing procedures are under implementation.

[2] A probability level is calculated, for example, on the basis of historical rates of returns and risk level (e.g. by using Value at Risk and Sharpe ratio measures).

[3] For example, we want to invest 1000 EUR, but at the price we can pay up only securities in the amount of 340 EUR are available other securities are too expensive in order to achieve the positive rate of return.

References

[1] O. Badawy and A. Almotwaly, "Combining neural network knowledge in a mobile collaborating multi-agent system", *Electrical, Electronic and Computer Engineering,. ICEEC '04, International* Conference on,2004, pp.325,328, doi: 10.1109/ICEEC.2004.1374457

[2] A. Bytniewski (ed.), *Architecture of integrated management information systems*, Wroclaw University of Economics Press, Wroclaw 2005.

[3] A. Bytniewski, A. Chojnacka-Komorowska, M. Hernes and K. Matouk, "The Implementation of the Perceptual Memory of Cognitive Agents in Integrated Management Information System", in: D. Barbucha, N. T. Nguyen, J. Batubara, *New Trends in Intelligent Information and Database Systems*, Studies in Computational Intelligence Volume 598, Springer International Publishing Switzerland, 2015, pp 281-290. doi: 10.1007/978-3-319-16211-9_29

[4] Cognitive Computing Research Group, http://ccrg.cs.memphis.edu/, data odczytu: 29.10.2014.

[5] W. Duch, „Sztuczna Inteligencja Reprezentacja wiedzy II: sieci semantyczne", https://www.fizyka.umk.pl/~duch/Wyklady/AI/AI06-1.ppt [02.05.2015].

[6] W. Duch, Architektury kognitywne, czyli jak zbudować sztuczny umysł, in: R. Tadeusiewicz (ed.) Neurocybernetyka teoretyczna, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2010.

[7] S. Franklin, F. G. Patterson, "The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent", in: *Proc. of the Int. Conf. on Integrated Design and Process Technology*, San Diego, CA: Society for Design and Process Science, 2006.

[8] R. Fikes and T. Kehler., "The role of frame-based representation in reasoning". *Commun. ACM* 28(9), 1985, pp. 904-920. DOI=10.1145/4284.4285.

[9] D. Fensel, *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag, New York, 2001.

[10] T.R. Gruber, "A Translation Approach to Portable Ontology Specifications", *Knowledge System Laboratory*, Academic Press Stanford University, 1993.

[11] M. Hernes, "A Cognitive Integrated Management Support System for enterprises", in D. Hwang, J. Jung, N.T. Nguyen (eds.), *Computational Collective Intelligence Technologies and Applications*, Lecture Notes in Artificial Intelligence, vol. 8733, Springer-Verlag, 2014, pp. 252-261. doi: 10.1007/978-3-319-11289-3_26

[12] M. Hernes and N.T. Nguyen, "Deriving Consensus for Hierarchical Incomplete Ordered Partitions and Coverings", *Journal of Universal Computer Science* 13(2)/2007, pp. 317-328.

[13] M. Hernes and J. Sobieska-Karpińska , "Application of the consensus method in a multiagent financial decision support system", *Information Systems and e-Business Management*, Springer Berlin Heidelberg 2015, doi: 10.1007/s10257-015-0280-9.

[14] D. Hofstadter and M. Mitchell, "The copycat project: A model of mental fluidity and analogy-making", in D. Hofstadter and the Fluid Analogies Research group, *Fluid Concepts and Creative Analogies. Basic Books*. Chapter 5, 1995.

[15] R. Katarzyniak, *Gruntowanie modalnego języka komunikacji w systemach agentowych*, Akademicka Oficyna Wydawnicza EXIT, 2007.

[16] M. A. Kadhim, A. Alam and M. K. Harleen, "A Multi-intelligent Agent Architecture for Knowledge Extraction: Novel Approaches for Automatic Production Rules Extraction*", International Journal of Multimedia & Ubiquitous Engineering*; Vol. 9 Issue 2, 2014, p.95.

[17] J. Korczak, H. Dudycz and M. Dyczkowski, "Design of Financial Knowledge in Dashboard for SME Managers", in: *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 1111-1118.

[18] J. Korczak and F. Hammadi-Mesmoudi, "A way to improve an architecture of neural network classifier for remote sensing application", Neural Processing Letters 1(1), pp.13-16, 1994.

[19] S. P. Li, Q. W. Yin, Y. J. Hu et al., "Overview of researches on ontology" *Journal of Computer Research and Development,* 2004, 41(7), pp.1041-1052.
[20] M. Maleszka and N.T. Nguyen, "Integration computing and collective intelligence", *Expert Systems with Applications,* vol. 42 (1), , 2015, pp. 358-378. doi:10.1016/j.eswa.2014.07.036
[21] N. T. Nguyen," Processing inconsistency of knowledge in determining knowledge of collective", Cybernetics and Systems: An International Journal, 40 (8), 2009, pp. 670-688.
[22] F. Pulvermuller, *The Neuroscience of Language. On Brain Circuits of Words and Serial Order,* Cambridge University Press 2003.
[23] X. Z. Wang, S. F. An., "Research on learning weights of fuzzy production rules based on maximum fuzzy entropy", *Journal of Computer Research and Development,* 43(4),2006, pp.673-678. doi: 10.1360/crad20060416
[24] J. Snaider, R. McCall and S. Franklin, "The LIDA Framework as a General Tool for AGI" *The Fourth Conference on Artificial General Intelligence,* 2011. doi: 10.1007/978-3-642-22887-2_14
[25] J. Sobieska-Karpińska and M. Hernes," Consensus determining algorithm in multiagent decision support system with taking into consideration improving agent's knowledge", in: *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems,* Wrocław 2012, pp. 1035-1040.
[26] J. F. Sowa, *Semantic Networks,* http://www.jfsowa.com/pubs/semnet.htm [02.05.2014].
[27] Z. Zeng, "Construction of knowledge service system based on semantic web", *Journal of The China Society For Scientific and Technical Information,* 24(3), 2005, pp.336-340.
[28] G. J. Zhu and Y. M. Xia, "Research and practice of frame knowledge representation", *Journal of Yunnan University (Natural Sciences Edition),* 28(S1), 2006, pp.154-157.
[29] T. Atanasova, T., "Towards semantic-based process-oriented control in digital home", *Federated Conference on Computer Science and Information Systems (FedCSIS), 2014,* pp.1133-1137, doi: 10.15439/2014F317.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 203–207
DOI: 10.15439/2015F324
ACSIS, Vol. 6

Smart Services Classification Framework

Tatiana Gavrilova, Liudmila Kokoulina
Graduate School of Management,
St. Petersburg University,
Volkhovsky per. 1-3, St. Petersburg, Russia

Email: {gavrilova, l.kokoulina}@gsom.pu.ru

Abstract— **The main goal of the study is development of the classification framework of smart service attributes as a first step in developing methodology of smart services implementation for Enterprise Information Portal (EIP) maintenance. First, we analyze available definitions of the "smart services" concept and concepts related to it: smart services are based on the idea of co-creation of value and rely on machine intelligence in connected systems. Second, we describe attributes of EIP services. Finally, we propose a new extended approach of the smart service attributes classification based on the list of characteristics of the EIP services. Our results contribute to the field of smart service research as well as to EIP-related studies both for academics and practitioners, as the proposed classification framework could serve as a basis for creation of smart services typology for the purpose of EIP maintenance.**

I. Introduction

THE concept of smart services has been evolving for decades, however, the field of smart services is still appears to be under development [1]. Publications devoted to smart services usually analyze this phenomenon from the practical perspective. Still, there is a need of further theorizing and conceptualizing of smart service concept.

One of the promising directions of smart services implementation is creation of the smart service system for the Enterprise Information Portal support. Enterprise Information Portal (EIP) serves as a unique point of contact for users providing information and supporting business decisions. EIP maintenance includes such processes as knowledge elimination, new knowledge regulation, and support. Proper EIP maintenance requires highly skilled professionals, and not all organizations potentially interested in EIP implementation would therefore agree to this endeavor.

As a part of our project "Ontology-based INtelligent Services for the knowledge PORTals support (InS-PORT)" we develop a methodology of creation of smart service system which would help to maintain EIP. The purposes of this smart service system are to support knowledge base formation, to eliminate outdate or improper information, to support the process of new knowledge regulation. However, this task requires deep understanding of smart service capabilities and structure. To the authors` knowledge, there are no developed smart service classifications. As a first step in dealing with this problem, we developed a smart service attributes` classification which is present in this short paper.

Our study is important for the smart service research stream due to several reasons. First, classification of the emerging phenomena enhances a uniform and standardised terminology. Besides, the classification helps to understand the definition of the object of classification.

This paper is structured as follows. In the next section, theoretical background of the smart service phenomena is presented, the various definitions are discussed, and the purpose of classification is justified. In the subsequent section, the research methodology is described and classification scheme procedural model is given. Next section of the article introduces the research on Enterprise Information Portal. As the result of the study, the developed smart service attributes classification framework is described in the subsequent section. The paper ends with conclusion remarks, limitations and future research.

II. Review of the Literature

First, we analyze available definitions of the "smart services" concept and concepts related to it. Currently, there are very few peer-reviewed publications on this topic, and most of them do not provide any formal definitions. The majority of definitions are very general or ambiguous. For example, one of the top experts in the field, Paul P. Maglio introduces smart services as:

"… capable of self-detection, self-diagnostic, self-corrective, or self-controlled functions through the incorporation of technologies for sensing, actuation, coordination, communication, control, and more" [2].

Some authors define smart services through description of their distinctive characteristics:

"Smart services are a wholly different animal from the service offerings of the past. To begin with, they are fundamentally preemptive rather than reactive or even proactive. Preemptive means your actions are based upon hard field intelligence; you launch a preemptive strike to head off an undesirable event when you have real-world evidence that the event is in the offing" [3].

Some authors claim that the use of term is often

speculative, and that smart services are simply "…a marketing term to bring together various meanings of the term Service (economic, technical, political, business- and end user- oriented) with an adjective to make it sound clever" [4].

The term "service" is used here as "…a function of an enterprise that is exposed through various technology-supported channels, and is amenable to re-use and composition into larger services which add value" [4]. It is important to mention that recently a new research stream appeared, labelled as "service science". Service science takes most of its inspiration in recent IT services growth and has been actively supported by IBM. Scholars in this field are still providing rather general definitions of term "services" such as "…as clients and providers working together to transform some state, such as material goods, information goods, organizations, which bears some ownership relation to the client" [5]. The two main issues that are recognized as basic tenets of the service science are: (1) co-creation of value by producer and client and (2) broad implementation of information technology [2].

Furthermore, the term "smart" implies two main properties. First, it highlights anthropomorphic features of the smart service. For example, technology research company Gartner, Inc. claims that smart technologies are "… technologies that do what we thought only people could do. Do what we thought machines couldn't do" [6]. Second, term "smart" is usually related to artificial intelligence (i.e. intelligence of machine) "[…] because it is impractical to deploy humans to gather and analyze the real-time field data required, smart services depend on "machine intelligence" [3].

In summary, this short literature review demonstrates that there is no agreement on what "smart services" are. However, based on several streams of thought, we can identify some key elements which are common in most definitions and which can help to come up with the working definition. Those key elements are 1) machine intelligence, 2) connectedness and 3) value co-creation by client and provider of a service. Thus, smart services are based on the idea of co-creation of value and rely on machine intelligence in connected systems [1]. Speaking of the attempts of smart services classification, there are no visible papers on this subject.

III. Research methodology

According to [7], 'a classification scheme consists of a set of characteristics which are suitable to classify objects of a specific application domain'. We have chosen a characteristic-based classification principle to develop a classification framework for smart service attributes since we suppose that the classification criteria might not necessarily be mutually disjunctive.

In our research, we follow the classification methodology proposed by [7]. This methodology suggests six general guidelines: completeness (the domain should be entirely covered by the classification scheme), precision (measure of detailing), consistency (lack of contradictions), extensibility (possibility to add or remove classes), user-friendliness (measure of how clear and understandable is the classification system), economic efficiency (related to costs associated with classification system implementation).

The procedural model to develop a classification scheme, described in [7], contains five phases: inception, elaborate characteristics, specify classification scheme, test, and use and maintenance. The first two stages include defining goals, conducting literature search in order to acquire a comprehensive set of potential characteristics. In our case, the literature concerning the EIP suggested the list of potential smart service classification characteristics. On the third stage the specification of the classification scheme was made, including defining principle of classification, selection and explanation of relevant characteristics, and defining relations between characteristics of the classification scheme. The last two phases ('test' and 'use and maintenance') are necessary for justification of the proposed classification framework, however, they are out of the scope of this paper and require further research.

IV. The Enterprise Information Portal

As the main goal of our research is to create methodology for smart services implementation for the purpose of EIP maintenance, a review of theory behind both concepts is required. Therefore, we analyze the research stream devoted to EIP, smart services, and their communalities, in particular how smart services correlate with EIP services.

Smart services as a subject of studies lies on the intersection of the scientific and technological paradigms of the information systems, knowledge management systems (KMS), enterprise information portals, service systems and smart services.

At the roots of the artificial intelligence studies there was a concept of "knowledge-based system" (KBS) [8], while the notion of knowledge management system (KMS) appeared much later in the management literature, and it is much wider than KBS. KMS include methods and techniques for the search, analyses, structuring, systematization, update and distribution of the information [9].

The term "enterprise information portal" was introduced in 1998. EIP is comprised of applications allowing companies to disclose information stored internally and externally, and to give the users the unique point of access and personalized information necessary for the decision making process in business [10]. The body of literature of this subject distinguishes two types of enterprise information portals: enterprise information portals and enterprise knowledge portals. The former type includes portals with services of search, exchange and sharing of the information. The latter type includes services developed with artificial intelligence methods. For our purposes, we define knowledge portals as the systems of knowledge management with the system of access embedded through enterprise portal.

While considering only technological component of the service systems (which is a composition of the interconnected information systems) a property of intelligence is identified. The property of intelligence is achieved by knowledge base inclusion and/or context awareness obtained by sensors, dynamic scalability, etc. [11; 12]. This type of intelligence is closely related to big data analytics. Recently scholars have argued that as more software and embedded intelligence is integrated in industrial products and systems, predictive

technologies based on big data will be used to predict product performance degradation, and autonomously manage and optimize product service needs [13].

Smart service systems could be considered as a sub-category of the intelligent systems. Smart service systems often have the following characteristics of the intelligent system:

- Self-configuration (or at least easy-triggered reconfiguration) [14; 15],
- Proactive behaviour (capability for prognosis or preventive actions, as opposed to the reactive behaviour) [3],
- Interconnectedness and continuous interactivity with internal and external system elements [16].

However, there are no commonly accepted definitions of intelligent and smart services – these terms are still developing [3].

EIP structures are based on the service-oriented architecture where services are located in the separate module. Basic portals` services include information search and exchange, communication among users, collaborative usage of the information. The technical services, which support EIP, are presented in the table I.

TABLE I.
ENTERPRISE INFORMATION PORTAL SERVICES

	Services	Functions
Basic	Communicational	Information exchange, collaboration between users and portal`s technical support group, realization of the modern voting and survey tools
	Informational	Notification of users about changes of events in their spheres of interests
	Navigational	Information search, search efficiency optimization
	Analysis and visualization of the spatial data	Thematic search services, services of the analysis and visualization of the spatial data (GIS portals)
	Personalized/identification	Identification, authorization and authentication of the portal`s users, portal visualization adaptation based on the user`s preferences (e.g., personal "cabinet" on the portal which stores the user`s profile and preferred system settings)
	Educational	Education of the employees
Technical	Statistical	Collection and analysis of the statistical information accumulated in the portal
	Audit	Logging of all actions included in the security system list
	Monitoring	Monitoring service

V. SMART SERVICES ATTRIBUTES CLASSIFICATION

Based on the review of smart services capable to enhance EIP maintenance, the following main attributes are found:

- Types of the elements comprising the service [3],
- Structure of the interactions among different types of the elements comprising the service [2; 3; 16; 17],
- The level of "intellectuality" or "intelligence" of the service [11; 12; 14],
- Dynamic aspects of the service working process [3; 16],
- Types of the information available to the service [8; 18], etc.
- Physical realization of the service (Software-as-a-Service, Hybrid cloud, own servers of the organization).

Following the classification methodology proposed in [7], we divided these attributes into two sub-groups: basic or IT implemented, and abstract (not dependent on IT implementation).

Basic attributes reflect physical implementation of the smart services including organization of the elements and IT platform.

Abstract attributes point our actual functionality of the service important for the business goals achievement. Those include dynamic properties, degree of knowledge awareness, and type of intelligence. The description of the service types is provided in the table II. The final classification of the smart services is presented in Fig. 1.

TABLE II.
SMART SERVICE ATTRIBUTES

Smart service attributes			Comment
Abstract, or modelling related	Dynamic properties	Without modelling of the changing environment; Past-based modelling; Stochastic modelling.	Modelling of the changing environment could be based on the analysis of the past or the probabilistic estimation of events.
	Intelligence	Knowledge-based; Data-based; Content-based.	Intelligence engine embedded in the smart service could be based on content (letters, audio-, video-, etc.), data (facts and features gathered from observation, measurement, sensors, etc.), and knowledge (rules and principles obtained from experience or theory).
	Knowledge awareness	Context-oriented; Explicit knowledge; Business intelligence.	Smart services could be based on knowledge derived from context (related to the user, environment, situation [19]), explicit knowledge (archived documents, charts), and business intelligence (OLAP and decision support systems).
Basic, or IT implemented	IT platform	Mobile; SaaS; Hybrid cloud; Corporate servers.	The choice of IT platform depends on the goals of the smart service: for internal purposes requiring confidentiality corporate servers may be used, however, SaaS and mobile platforms are gaining popularity with growing reliability and security.
	Elements	IT; People; Hybrid.	As smart service usually is a socio-technical system, the elements comprising it could be both IT (user- or network- oriented) and people (users, analysts, developers, support team).

In order to create a visualized from of smart service attributes classification, we illustrate our results with the mind map presented in fig. 1.

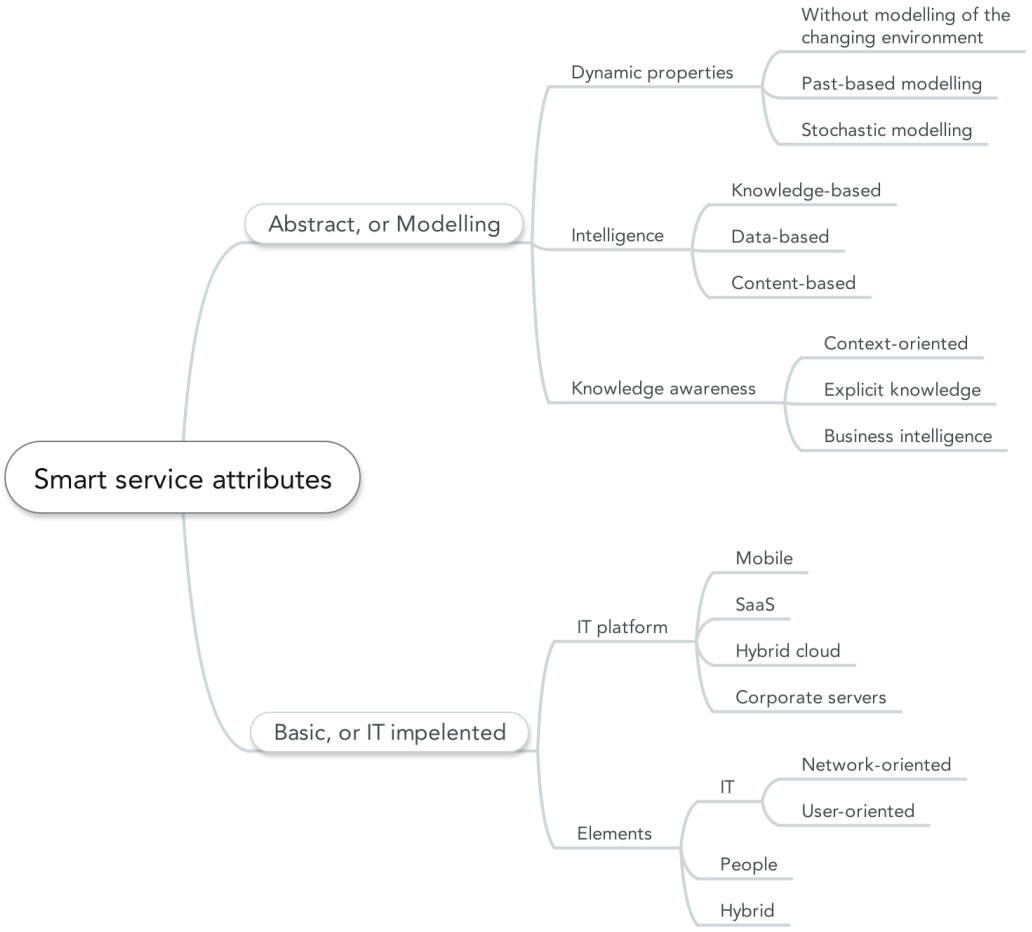


Fig. 1. Smart Services Attributes Classification

The proposed classification framework serves as a starting point in developing methodology of smart services implementation for the purpose of EIP maintenance. However, this classification could be generalized to other cases of smart services implementation. Therefore, our results contribute to the theory behind smart service systems. Moreover, our classification will be helpful to practitioners interested in smart services implementation.

VI. Conclusion

The main results of this study can be summarized as following. First, smart services are a relatively new concept that emerged because of progress in machine intelligence, global connectivity and big data. Second, the smart service system could be analyzed through the lenses of established knowledge management methods. The main contribution from this perspective is the development of new smart service attributes classification based on the characteristics derived from Enterprise Information Portal services analysis.

As of limitations of this research, the classification scheme creation procedure requires to test and revise the proposed scheme, therefore, more work is needed in order to test whether different smart services could be placed into it, to measure quality criteria (relevance, completeness), and to generate statistics of attributes` use.

Further research is required for the development of smart service typology and decision tree related to smart services implementation for the purpose of EIP maintenance and other contexts.

Acknowledgment

The research is supported by the grants of Russian Science Foundation (project No. 15-18-30048) and Russian Foundation for Basic Research (project No. 14-07-00294).

References

[1] S. Vlasov, T. Gavrilova, "Smart services: state-of-the-art", in *On-line proceedings of International conference "GSOM Emerging Markets Conference: Business and Government Perspectives"*, St. Petersburg, 2014, pp. 523-530.

[2] P. P. Maglio, "Editorial column – Smart Service Systems", *Service Science*, vol. 6, no. 1, 2014, pp. 13–15. http://dx.doi.org/10.1287/serv.2014.0065

[3] G. Allmendinger, R. Lombreglia, "Four strategies for the age of smart services", *Harvard Business Review*, Sep. 2005, pp. 1–11.

[4] K. Duddy, "What Would Smart Services Look Like",in G. Feuerlicht & W. Lamersdorf (Eds.), *Service-Oriented Computing – ICSOC 2008 Workshops*, vol. 5472, pp. 5–14. http://dx.doi.org/10.1007/978-3-642-01247-1_2

[5] J. Spohrer, P. P. Maglio, "The Emergence of Service Science: Toward Systematic Service Innovations to Accelerate Co-Creation of Value", *Production and Operations Management*, vol. 17, no. 3, 2009, pp. 238–246. http://dx.doi.org/10.3401/poms.1080.0027

[6] T. Austin, "The Disruptive Era of Smart Machines Is Upon Us", Gartner, Inc., 2009.

[7] P. Fettke, P. Loos, "Classification of reference models: a methodology and its application", *Information Systems and e-Business Management*, vol. 1, 2003, pp. 35-53. http://dx.doi.org/10.1007/BF02683509

[8] L. Shu-Hsien, "Expert system methodologies and applications—a decade review from 1995 to 2004", *Expert Systems with Applications* vol. 28, no. 1, 2005, pp. 93-103. http://dx.doi.org/10.1016/j.eswa.2004.08.003

[9] D. G. Schwartz, *Encyclopedia of knowledge management*, IGI Global, 2006.

[10] C. C. Shilakes and J. Tylman, *Enterprise Information Portals*, NY, Merrill Lynch, 1998.

[11] H. A. Simon, *The sciences of the artificial*, MIT Press, 1996.

[12] R. A. Brooks, "Intelligence without representation", *Artificial Intelligence* vol. 47, no. 1, 1991, pp. 139-159. http://dx.doi.org/10.1016/0004-3702(91)90053-M

[13] J. Lee, H. Kao, S. Yang. "Service innovation and smart analytics for Industry 4.0 and big data environment", in *Proc. of the 6th CIRP Conference on Industrial Product-Service Systems*, 2014. http://dx.doi.org/10.1016/j.procir.2014.02.001

[14] S. Barile, F. Polese, "Smart service systems and viable service systems: Applying systems theory to service science", *Service Science* vol. 2, no. 1, 2010, pp. 21-40. http://dx.doi.org/10.1287/serv.2.1_2.21

[15] D. Tranfield, D. Denyer, P. Smart, "Towards a methodology for developing evidence-informed management knowledge by means of systematic review", *British Journal of Management*, vol. 14, no. 3, 2003, pp. 207-222. http://dx.doi.org/10.1111/1467-8551.00375

[16] N. Gershenfeld, R. Krikorian, D. Cohen, "The Internet of things", *Scientific American, vol.* 291, no. 4, 2004, p. 76. http://dx.doi.org/10.1038/scientificamerican1004-76

[17] A. Fano, A. Gershman, "The future of business services in the age of ubiquitous computing", *Communications of the ACM* vol. 45, no. 12, 2002, pp. 83-87. http://dx.doi.org/10.1145/585597.585620

[18] M. Alavi, D. Leidner, "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues", *MIS Quarterly*, vol. 25, no. 1, 2001, pp. 107-136. http://dx.doi.org/10.2307/3250961

[19] E. Pascalau, G. J. Nalepa, and K. Kluza, "Towards a Better Understanding of Context-Aware Applications", in *Proc. of the 2013 Federated Conference on Computer Science and Information Systems*, pp. 959–962.

10th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

Topics

The areas and topics of interest include, but are not limited to two groups:

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles

Event Chairs

Arogyaswami, Bernard, Le Moyne University, USA
Chmielarz, Witold, University of Warsaw, Poland
Karagiannis, Dimitris, University of Vienna, Austria
Kisielnicki, Jerzy, University of Warsaw, Poland
Ziemba, Ewa, University of Economics in Katowice, Poland

Program Committee

Antosova, Maria, Technical University of Košice
Bialas, Andrzej, Institute of Innovative Technologies EMAG, Poland
Christozov, Dimitar, American University in Bulgaria, Bulgaria
Csikosova, Adriana, The Technical University of Košice, Slovakia
DeLorenzo, Gary, California University of Pennsylvania, United States
Dima, Ioan Constantin
Dudycz, Helena, Wrocław University of Economics, Poland
Espinosa, Susana de Juana, University of Alicante, Spain
Gafni, Ruti, The Academic College Tel-Aviv-Yaffo, Israel
Geri, Nitza, The Open University of Israel, Israel
Grabara, Janusz, Czestochowa University of Technology, Poland
Jelonek, Dorota, Czestochowa University of Technology, Poland
Kersten, Grzegorz, Concordia University, Montreal, Poland
Kobyliński, Andrzej, Warsaw School of Economics, Poland
Kohun, Frederick, Robert Morris University, United States
Korczak, Jerzy, Wrocław University of Economics, Poland
Lasek, Mirosława, University of Warsaw, Poland
Levy, Yair, Nova Southeastern University - Graduate School of Computer and Information Sciences (GSCIS), United States
Miliszewska, Iwona, University of Canberra, Australia
Modrak, Vladimir, The Technical University of Košice, Slovakia
Niedźwiedziński, Marian, University of Lodz, Poland
Owoc, Mieczyslaw, Wroclaw University of Economics, Poland
Pańkowska, Małgorzata, University of Economics in Katowice, Poland
Pastuszak, Zbigniew, Maria Curie-SKlodowska University, Poland
Phusavat, Kongkiti, Kasetsart University in Bangkok, Thailand

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 211–216
DOI: 10.15439/2015F21
ACSIS, Vol. 6

ITGovA: Proposition of an IT governance Approach

Adam CHEKLI
Hassan II University - Mohammedia
Ben M'sik Faculty
Email: adamchakli@gmail.com

Sara AREZKI
Hassan II University - Mohammedia
Ben M'sik Faculty
Email: sara.arezki@gmail.com

Abdelouahed NAMIR
Hassan II University - Mohammedia
Ben M'sik Faculty
a.namir@yahoo.fr

***Abstract*—I To cope with issues related to optimization, rationalization, risk management, economic value of technology and information assets, the implementation of appropriate IT governance seems an important need in public and private organizations. It's one of these concepts that suddenly emerged and became an important issue in the information technology area. Many organizations started with the implementation of IT governance to achieve a better alignment between business and IT, however, there is no method to apply the IT governance principles in companies. This paper proposes a new approach to implement IT governance based on five iterative phases. This approach is a critical business process that ensures that the business meets its strategic objectives and depending on IT resources for execution.**

***Keywords*—Information technology, IT governance, lifecycle, strategic alignment, value**

I. INTRODUCTION

To get more competitive and to face an increasingly fierce competition, businesses restructure to streamline operations and jointly benefit from advances of information technology to improve their competitiveness and align with other companies. Focusing competitiveness and cost value push companies to trust information technology witch become an essential pillar of the most company's strategy... However, to realize that IT projects can return value to the organization and increase its performance unless they may include various ethical risks can explain the relatively attraction for the IT governance.

IT governance is a mechanism to meet new challenges of IT resources management policy. It determines the rules, procedures, structures and behaviours, to a better relationship between the involved actors in the operation and information system administration within an organization. For many, the prospects that propose this approach are synonymous higher operating efficient systems designed to free up additional optimum performance capacity [2].

This paper proposes an approach to improve IT governance. This approach is based on phases allowing an initial assessment of company's IT governance maturity, followed by a definition of a realistic target to reach. The selected target will be translated into best practices actions plan composed of a hierarchical prior priority to improve IT governance processes. Like any project, performance evaluation indicators will be established and accompany the project life cycle.

II. IT GOVERNANCE

Governance is a key concept for the information system and information technology. Today, IT can answer many crucial questions in this area [3]:

- How to manage the relationship between top management and IT department?
- What are roles and responsibility limits between the directions witch use and manage the information of the company.
- How to improve information system efficiency?
- What are the key processes of IT department?
- How to manage and organize IT department?
- How to ensure a sustainable information system?

The emergence of IT governance issue is associated with the development of the following phenomenon [4]:

- The appearance of an enterprise IT department side of IT department branches.
- The questioning about IT value after a long period of IT investment.
- The rise of values such as transparency, accountability and precaution principle.

The ITGI [5] defines IT governance as « *an integral part of enterprise governance and consists of the leadership and organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategies and objectives* ». This means that the IT governance is very important to the enterprise and needs to be treated by top management.

Robert Roussey [6] describes IT governance as « *a term which is used to describe the way how those in charge of governance in an entity will consider it in the supervision, monitoring, control and management. The way how this governance is applied in an entity will have an immense impact on the achievement of its objectives, vision and strategic goals* ». This means that the IT governance is an

important way to improve business and to achieve enterprise objectives.

IT governance must be an integral part of the overall corporate governance. Its objective is to ensure that[7]:

- IT function is provided with the organization that supports it.
- IT function allows company to exploit the opportunities and maximize the value.
- IT resources are used in a reasonable and responsible way.
- IT risks are managed in an appropriate way.

According to the ITGI [4], the IT governance is a management process based on best practices allowing the company directing the IT functions in the goal to:

- Support its objectives of creating value
- Improve the performance of IT processes
- Master the financial aspects of IT
- Develop IT solutions and skills that the company will need in the future.
- Ensure that the IT risks are managed
- While developing transparency.

After this brief presentation of IT governance and its concepts, the paper will detail the different phases of the approach lifecycle.

III. Presentation of the proposed IT governance approach

This section will provide an overview of the proposed methodological approach of IT governance.

This method is based on IT governance concepts defined by the ITGI, best practice frameworks and feedbacks of companies that have already implemented an IT governance plan.

The methodological approach is in the form of phases declined in activities. Each phase is described by a sheet that has the following items:

- Phase ranking in the process
- Phase description
- Phase objectives
- Input elements and deliverables
- Satisfaction condition
- Needed roles to achieve the objectives
- Major risks to consider
- Human and technical required skills
- Tools
- Presentation of the activities constituting the phase
- Remarks

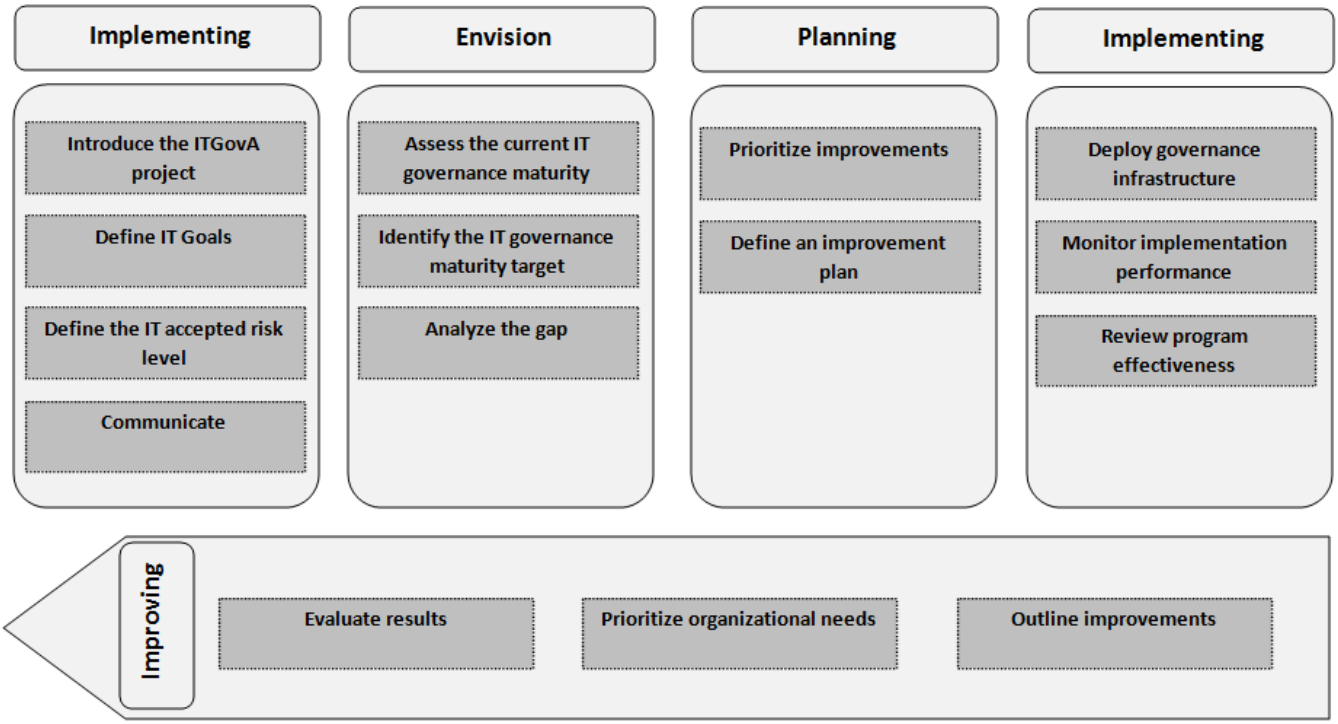
Each activity is described by a sheet that includes the following items:

- Activity objectives
- Activity description
- Actions to do
- Input elements and deliverables
- Responsibilities
- Remarks

The proposed approach is not rigid. This is a set of methodological components built to be selected and integrated in order to form the best solution to a specific need in a specific situation. The strength of this approach is given in the way how its components are designed corresponding to the most common project management sequence. Figure 1 presents the phases of the proposed approach and its activities, roles and deliverables.

FIGURE I.

THE PROPOSED IT GOVERNANCE APPROACH



IV. THE PROPOSED IT GOVERNANCE APPROACH LIFECYCLE

A. *Inception*

The inception phase aims to understand the project scope and objectives and getting enough information to confirm that the project should proceed. In this phase, significant business and requirements risks must be addressed.

The objectives of the inception phase include:

- Establish the project scope and boundary conditions
- Obtain the agreement of stakeholders about the scope of the project
- Define the initial program business case
- Obtain the acceptance of the stakeholders about the initial cost and schedule estimates.
- Indentifying, assessing, and accepting risk level
- Communicate about the IT governance project.

Define resources from the sponsor to all the stakeholders. Envisage any intern or extern training.

If the stakeholders agree that the project meets the above criteria, the project move to the Envision phase. If the project fails in any area above, the project may be re-directed or cancelled outright.

Table 1 describes the activities of the Inception phase.

TABLE 1. THE INCEPTION PHASE ACTIVITIES

Activities	Description
Introduce the IT governance project	This activity of introducing the IT governance project aims to obtain an understanding of the IT governance program, background and objectives. Define the initial program concept business cases and to obtain the buy-in and the commitment of all the stakeholders.
Define It goals	The activity of defining goals aims to define IT goals based on business goals for IT while considering current and required future service and the enterprise architecture for IT
Define the IT accepted risk level	This activity of defining the IT accepted risk level aims to obtain an understanding of the enterprise present and future attitude toward risk and how it will impact the project
Communicate	This activity aims to ensure that all parties are involved. Committed and knowledgeable about the objectives of the project

B. *Envision*

Before implementing the IT governance program, a good assessment of IT governance maturity is needed. This gives all stakeholders a clear view of the current state. After evaluating this current state, a definition of a target maturity of each IT governance process is needed followed by a gap analysis between the current and the target state allowing a translation of the differences into improvement opportunities.

Table 2 describes the activities of the Envision phase.

TABLE 2. THE ENVISION PHASE ACTIVITIES

Activities	Description
Assess the current IT governance maturity	The activity of assessing the current IT governance maturity should determine the current capability maturity of all processes. This assessment should be based on questionnaires, interviews and studies of existing documents.
Identify the IT governance maturity target	This activity aims to define a capability maturity level to achieve. This identification can consider market best practices already done by other companies in the same sector.
Analyze the gap	This activity aim to analyze the difference between the two states (current and future) of maturity. It's essential to transform differences in good opportunities.

C. *Planning*

The Planning phase is the third phase in the approach lifecycle. It involves creating a set of plans to guide the project team through the execution and closure phases of the project. Plans help to manage time, cost, quality, change, risks and issues. They will also help manage staff and external suppliers, to ensure that you deliver the project on time and within budget.

The objectives of the Planning phase are:

- Translate improvement opportunities into justifiable projects
- Prioritize and focus in the high impact projects
- Integrate the improvement projects into the overall program plan

Table 3 describes the activities of the Planning phase.

TABLE 3. THE PLANNING PHASE ACTIVITIES

Activities	Description
Prioritize improvements	This activities aims to translate improvement opportunities into justifiable projects. Prioritize and focus on high impact projects.
Define an improvement plan	This activities aims to integrate the improvement opportunities into the overall grogram plan

D. *Implementing*

The Implementing phase concerns the design and the deployment of the IT governance solution. The Implement phase is decomposed into two parts: deploying and monitoring.

The implementing phase has the following objectives:

- Establish a set of requirements for the design of the governed process
- Design and document the governed process including:
    - Role responsibilities for the execution of specific governed processes
    - Detailed descriptions of the activities required in support of the governed process.
- Deploy the governed process
- Ensure the successful deployment of the operationlized governance solution
- Transfer skills to practionners to manifest organizational change
- Audit skills to practionners to manifest organizational change
- Manage exceptions to processing based upon the needs of the business.

Table 4 describes the activities of the Implementing phase.

TABLE 4. THE IMPLEMENTING PHASE ACTIVITIES

Activities	Description
Deploy governance infrastructure	This activity of deploying governance infrastucture Implement the detailed improvement project, leveraging enterprise program and project management capabilities, standards ad practices
Monitor implementation performance	This activity aims to Integrate the metrics for project performance and benefits realization of the governance improvement project into the performance measurement system for regular and ongoing monitoring
Review program effectiveness	The activity of reviewing programm effectiveness Assesses the result and experience gained from the program. Record and share any lessons learned

E. *Improving*

The Improving phase provides formal feedback to the board of directors, lines of business, and projects, based

upon the predefined goals that are substantiated in the business case of governance. The results are evaluated based upon the key performance (KPIs) and key goal indicators (KGIs) that are defined during the inception phase.

The objectives of the improvement phase are:

- Evaluate the performance criteria of your IT governance solution against its fulfillment of the
  - The measures or metrics that are used for control of the governance process
  - The measures or metrics that are used for the control of the management process
- Prioritize the current organization needs as a function of updating the current measures or process of either the management or governance functions.
- Identify and provide the business justification, outlining the potential improvements that can be realized by altering the existing processes

Table 5 describes the activities of the Improving phase.

TABLE 5. THE IMPROVING PHASE ACTIVITIES

Activities	Description
Evaluate results	This activity engages the Governance board in the examination of the baggregated metrics from source measures that were based upon the KPIs and KGIs that were defined during the Implementing phase. Evaluation of results is performed by comparing current results to a predefined baseline metric.
Prioritize organizational needs	The activity of prioritizing organization needs takes place when analysis of the operational and quality metrics is completed. This assessment activity critically evaluates the baseline metric to the gathered results for the identification of operational and quality improvements. It does this along with critical examination of the measures and metrics themselves to assess their applicability to the problem space that is being measured.
Outline improvements	This activity treates the critical task of documenting the objective data, and analysis results to specify suggested changes to the target systems. These changes are either within the governance process or to the management process that produces the product output. These suggested changes close the feedback loop on the governance process and the management process.

F. RECOMMENDATIONS

Implementing ITGovA approach requires discipline, committement and support of all stakeholders. IT

  goals and objectives set during the Plan and Implement phases.
- Evaluate product quality goals and objectives that are established during the Plan and Implement phases.
- Identify any required changes to the following items:
  - The governance process itself
  - The management process

governance should be adapted to any organizational structures, culture and overall business strategy.

Several success factors are emphasized for the success of the project:

- Committement of the top management: The support and commitment of project stakeholders is a cornerstone to its success. If no instructions are issued by the top management, the project will experience failure at any moment;

- Develop the organizational structure: The human capital is very important for the implementation of the process of IT governance, hence the importance of developing organizational governance structure first and also provide the required skills and seek external expertise to build teams;

- Develop processes: The prioritization process to implement must be well justified. This justification must take into consideration several factors such as the budget, the availability of resources; etc. Two options: Implement per-process or develop in lots of simultaneous processes. Use a method of iterative incremental development work;

- Don't start from crash: Use existing reference market. These standards are the result of a large expert and so they include a variety of good useful for the implementation of the approach practical reflection;

- Communication: The process of governance of information systems is new to the company. It must be well explained and continuously reinforced to win the commitment of all stakeholders. The approach should always be measured by performance indicators and targets are met. These measures should always be

updated to determine the impact of the process on the company.

G. CONCLUSION

This new approach is flexible and has an adaptable architecture that is designed to maximize the effectiveness of the developed IT governance solution. It describes what needs to be done in order to implement effective IT governance solutions.

This method highlights the importance of aligning the business, IT organizations and enterprise architecture to establish a basis from which strategic business value may be realized. The approach highlights also the relation between top management, stuff and auditors in planning, designing, implementing and evaluating the IT governance project.

H. REFERENCES

[1] Exler, R., 2004, IT Governance frameworks

[2] F.Georgel, 2009, IT governance: management stratégique des systèmes d'informations Dunod.

[3] ITGI. 2003, Board briefing on IT governance

[4] ITGI, 2007, IT governance roundtable: IT governance trends

[5] ITGI, 2008, IT governance roundtable: IT governance roundtable

[6] Longépé, C. 2006, Le projet d'urbanisation du SI. Dunod

[7] L.Muyller, M.Magee, P.Marounek, A.Philipson, 2008, IBM IT Governance Approach, IBM

[8] P.Weill, J.W. Ross, 2004, IT Governance. How Top Performers Manage IT Decision Rights for Superior Results. Harvard Business School Press

[9] P.Weill, Richard Woodham, 2002, Don't Just Lead, Govern: Implementing Effective IT Governance

[10] The national Computer Center, 2005, IT governance: Developing a successful governance strategy, National Computer Center

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 217–223

DOI: 10.15439/2015F61
ACSIS, Vol. 6

Case Based Reasoning as an Element of Case Processing in Adaptive Case Management Systems

Łukasz Osuszek
IBM Polska
Email:
lukasz.osuszek@pl.ibm.com

Stanisław Stanek
The General Tadeusz Kosciuszko Military
Academy of Land Forces, Poland
Email: s.stanek@wso.wroc.pl

***Abstract*— The paper sets out from a proposition that the concept of Case Based Reasoning could improve business decisions and optimize case processing in modern Adaptive Case Management (ACM) systems. While depicting the state of the art in the continued efforts to blend Artificial Intelligence (AI) with Business Process Management (BPM), Knowledge Management (KM) and Adaptive Case Management, the authors take notice of how the classical ACM platform has recently been evolving. The dynamic and adaptive nature of some business processes poses challenges that the classical BPM approach cannot adequately address. Adaptive Case Management has been developed to better cope with such challenges. ACM not only makes it easier to align a business to rapidly changing requirements and conditions, but it also enables organizations to more effectively exploit the potential inherent in the organizational knowledge and information resources. The paper discusses the evolution of ACM systems and proposes to apply Case Based Reasoning (naturally coupled with AI) in optimizing ACM outcomes.**

I. Introduction

A completely new approach employing the existing ACM tools and the Case Based Reasoning model has been developed to address the problem of decision support within business case processing. The idea involves integrating the Case Based Reasoning method into Adaptive Business Case Management.

This paper describes an approach that is founded on the application of Case Based Reasoning to deploy decision support and was preceded by a literature based discussion on Artificial Intelligence (i.e. the CBR method) and its application to support Adaptive Business Case Management. Attention is focused on ACM and CBR, with an aim to provide a complete theoretical framework for reflection on their proposed integration and for applying the integrated methods to exploring a set of business problem solutions.

The paper advances and investigates the following theses:

a) The Case Based Reasoning method may be used to support ACM by providing faster access to the information needed to make business decisions.
b) Supporting ACM with Case Based Reasoning makes the process of exploring a set of business problem solutions faster and more effective.

The paper deals with the application of the Case Based Reasoning method in supporting the ACM method. Case Based Reasoning (CBR) is an artificial intelligence method based on reusing the outcomes of previously solved problems: when a new problem arises, the problem solving process begins with an effort to find the closest matching solution to the problem within a set of historical solutions. Once a matching solution is found, it is adapted to the specific problem and an attempt is made to apply it. The new solution is too stored in a dedicated repository. With each subsequent problem solved, the repository becomes larger.

Adaptive case management processes are of dynamic character, since they are not defined until at runtime. To master the unpredictability of processes and hence facilitate process management in contexts where processes are mostly complex and where relevant decisions are affected by a large number of factors, more and more organizations choose to switch to Adaptive Process Management systems. ACM allows perfect visibility and full control of each specific case, whether it is handled by a predefined or an ad hoc process, or by a combination of the two.

An important part of a problem solving process is to define the case and represent it in a machine readable format, i.e. one that can be handled by a computer. If the case has been defined and represented accurately, and if the case repository is adequately structured, the process of recording a particular problem can be carried out parallel to problem solving.

The paper presents the authors' original solution that extends the capabilities of ACM through the addition of functionalities typical of CBR.

An enterprise that is run in line with the ACM concept will be intrinsically capable of combining its core business activities with a day-to-day ability to create and review innovative solutions. Since process operators can modify processes dynamically, the entire business management system is open to creative initiatives from staff at large, while at the same time avoiding chaos that might arise as a result of spontaneous changes to operating properties. In addition, since it possible to examine the outcomes of changes as they emerge, information on which practices and

solutions deliver the best results and which produce the worst can be appended to organizational collective knowledge. This stands for day-to-day improvements and adaptations to business processes relying on the best knowledge of a large portion of personnel and becoming validated through feedback from customers.

A fundamental principle of ACM is associated with the belief that any organization should continually collect, process and utilize knowledge on the mechanisms governing its business environment, and that such an approach is not only most effective, but simply essential if you want to be able to respond to customers' expectations and keep pace with the rapid changes in today's marketplace. ACM is often said to be focused on building a learning organization. Improvements to an organization's internal processes take place across several dimensions and engage executives and staff alike. The paper describes a methodology for integrating the AI-based CBR method into the ACM domain, thereby improving the standard ACM mechanisms.

It needs to be stressed that a dynamic Business Process Management model can be implemented within an organization irrespective of the products and/or services its offers, and that the effects of its implementation largely depend on the professional skill of personnel, their effectiveness in managing organizational knowledge, and their ability to make optimal business decisions – which entails the requirement for all staff to be involved in developing and formulating new solutions.

II. Background and rationale

The CBR field has been growing rapidly over the last 20 years, The increased interest in CBR is evident in the number of research papers presented at major conferences, and in the availability of commercial tools and successful applications in daily use.

What is Case Based Reasoning about? What it basically does is try to solve new problems by recalling previous instances of similar cases and reusing information and knowledge on those cases. CBR can be therefore described as a problem solving paradigm. It is able to utilize specific knowledge on previously experienced, concrete problem cases; it attempts to find a similar past case and reuse the related information in addressing a new problem. CBR is at the same time an approach allowing incremental, sustained learning, since new experiences are retained each time a problem has been solved, making them immediately available in addressing new problems that will arise in the future.

Under CBR terminology, a "case" usually denotes a problem situation. A previously experienced situation that which was captured and learned in a way that makes it possible to reuse the experience in solving future problems, is referred to as a "past case", "previous case", "stored case", or "retained case". Accordingly, terms such as "new case" or "unsolved case" refer to a description of a new problem to be solved. Case Based Reasoning is – in effect – a cyclical, integrative all-round process of problem solving, learning from this experience, and solving new problems.

The affinity between ACM and CBR goes beyond cycles alone. At the research level, ACM literature recommends that effective case management solutions target people, processes, information and technology [1]. From a CBR perspective, Aamodt and Nygård argued – decades ago – that CBR research should address practical applications and focus on optimizing the linkages between the CBR system and its user rather than on the CBR system alone [2]. This encourages a perception whereby CBR appears as an approach that contributes to ACM and Knowledge Management.

In the paper, it will be demonstrated that Case Based Reasoning (CBR) is intrinsically applicable to Case Management.

III. Learning from the CBR and ACM perspective

A very important feature of Case Based Reasoning is that it is intrinsically coupled with learning. The driving force behind Case Based methods stems, to a large extent, from machine learning (a subfield of machine learning [3]). Thus, no matter how cases are acquired, the notion of Case Based Reasoning designates a reasoning method as well as a machine learning paradigm that enables sustained learning by updating a case base on solving each problem. Under CBR, learning occurs naturally as a by-product of problem solving: when a problem is successfully tackled, relevant experience is retained with a view to solving similar problems in the future; if an attempt to solve a problem fails, the reason for the failure is diagnosed and memorized in order to avoid making the same mistake in the future.

Case Based Reasoning favors learning from experience, since it is usually easier to learn by retaining a concrete problem solving experience than to generalize from it. Still, effective learning in CBR requires sophisticated methods to extract relevant knowledge from experience, incorporate cases into an existing knowledge structure, and index them for subsequent matching with similar cases.

The term "memory" is often used to designate a storage structure that holds existing cases, i.e. a case base. Memory thus refers to what has been memorized from past experiences. Accordingly, "reminders" or "pointers" are structures referencing, or pointing to, some part of the memory.

ACM standardizes information, processes, and people, allowing for each case to be presented fully and in many aspects. Case management stands for coordinating the service activities undertaken in an effort to achieve a specific objective. Typically, it involves creating a case file and performing certain tasks (e.g. including the right documents in the file). These tasks may be standardized (or pre-defined) actions related to the type of process at hand, or actions that are designed and added *ad hoc* when dealing with

a particular case. Each case file contains a description of the customer, product, project, patient, etc. The description can be defined freely, either internally or externally.

What ACM essentially does is shift the process of knowledge gathering from the template analysis, modeling or simulation phase into the process execution phase. An ACM system collects actionable knowledge – without an intermediate analysis phase – from business users. All information that might be required in processing a case is stored in the ACM system (a repository, case history, case-related communications, etc.).

Moreover, ACM helps manage the unpredictable by enabling knowledge workers to effectively cooperate and share their knowledge, thus improving the functionality of any decision support system [4].

Users engaged in solving tactical and strategic problems will rather expect the system to become a "partner in problem solving." Interestingly enough, we have found that the lowest skill levels are associated with the highest expectations from the system, including a proactive attitude in assisting the user. Conversely, the expectations of most advanced and creative problem solvers are limited to being offered an efficient technology and a rich collection of presentation tools.

Knowledge workers will convert restricted-access knowledge into open-access knowledge, thus building up organizational resources of information/knowledge on business cases. The learning process results in expanding the organizational knowledge base and improving the staff's creativity, which allows business cases to be handled more effectively. Users can retain proven operating procedures within embedded structures and templates consisting of business case process components, such as data models, process models, user interface ingredients, rule sets, and case configurations. Users are allowed to add their own templates and create complete case processing applications that will satisfy their industry-specific needs and/or their specific expectations [5].

An effective Adaptive Case Management solution should be able to support organizational learning from previous cases. The learning may lead to defining new processes, designing new procedures, enhancing the efficiency of online help services, etc., whereby lessons learned by knowledge workers are immediately applied in the process improvement cycle. The phrase "formalized experience" is often used to describe established practices that have been transformed into automated steps and/or procedures aimed at assisting the processing of future cases.

The classical process improvement cycle (e.g. in BPM systems), administered by process leaders and involving such consecutive steps as process modeling, performance monitoring, formulating conclusions and, eventually, utilizing the findings to improve the process, is far too slow and therefore inadequate. What is more, in the event that some customers have conflicting expectations, it might be impossible to design a "universal" process that would be acceptable to all stakeholders.

The learning approach of Case Based Reasoning is sometimes referred to as Case Based learning. Central tasks that all Case Based Reasoning methods have to deal with are to identify the current problem situation, find a past case similar to the new one, use that case to suggest a solution to the current problem, evaluate the proposed solution, and update the system by learning from this experience.

IV. How CBR is related to ACM

At the highest level of generality, the following four processes may describe a general CBR cycle:

- RETRIEVE the most similar case or cases,
- REUSE the information and knowledge on a case to solve a current problem,
- REVISE the proposed solution,
- RETAIN those parts of the experience that are likely to be useful in future problem solving.

A new problem is solved by retrieving one or more previously experienced cases, reusing the case in one way or another, revising the solution based on reusing a previous case, and retaining the new experience by incorporating it into the existing knowledge base (case base). Each of the four processes involves a number of more specific steps that will be described in the task model.

The CBR paradigm comprises a range of methods for organizing, retrieving, utilizing and indexing knowledge retained from past cases. Cases may be kept as individual experiences or as generalized cases made up of sets of similar cases; stored as separate knowledge units, or split up into subunits and distributed within a knowledge structure; indexed by prefixed or open vocabulary, within a flat or a hierarchical index structure. Solutions from previous cases may be directly applied to current problems or modified to allow for differences. Matching the cases, adapting the solutions, and learning from experience may be guided and supported by either a deep model of general domain knowledge or by shallow and compiled domain knowledge; or else it may be based on apparent syntactic similarity alone. CBR methods may be wholly self-contained and automatic, or they may interact heavily with the user for support and guidance of their choices. Some CBR methods assume a rather large amount of widely distributed cases in their case bases, while others are based on a limited set of typical ones. Past cases may be retrieved and evaluated sequentially or in parallel [6].

CBR can be easily and effectively used in ACM environments. It can be seen as a natural extension to the existing ACM system functions and a way to improve organizational performance in solving business cases, managing organizational knowledge, and supporting decisions made by knowledge workers. With an ACM system in place, all the data and information related to each case, gathered while processing it, is stored in a case

repository. Descriptions contained in case files can be defined freely and may include e.g. customer and/or supplier information, applications or requests, customer information obtained from external sources, product specifications, financial reports, legal documents and opinions, statements, correspondence, test results, X-ray scans, photographs, technical drawings, and many other similar resources.

Case management typically involves creating a case file and performing certain tasks. These tasks may be standardized (pre-defined) or added dynamically when dealing with a case. All information pertinent to the case, accumulated in the case file, can be used by all personnel engaged in handling the case and working together in processing and closing it. Corporate executives can then use such information in monitoring the process of dealing with a particular case. This is an idea that is perfectly consistent with CBR.

All the steps and decisions taken by a knowledge worker, as well as all other information related to the processing of a particular case, are stored within ACM structures. The organization's internal, restricted-access expertise, guidance and resources needed to solve a specific case are all contained within ACM structures. Most frequently, the information on a case is structured as shown in the following diagram.

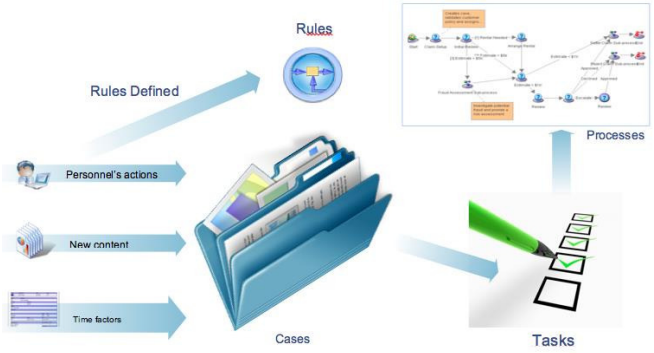


Fig. 1 A typical case structure within an ACM system

Once a case is closed, the information on the case is stored for audit purposes (in compliance with certain regulations) or for use in other long-term business processes. The idea described in this paper proposes to extend the application of such information to the processing of new cases via CBR.

ACM allows the employee to create rules by reference to a repository of previous cases representing best practices. The availability of information on similar problems, and on optimum solutions to these, leads to minimizing repetitive work.

V. Incorporating CBR into ACM

Under CBR methodology, an initial problem description defines a new case, which is then used to RETRIEVE a case from the collection of previous cases. The retrieved case is combined with the new case – through REUSE – into a solved case, i.e. a proposed solution to the initial problem. Through the REVISE process this solution is tested for success, e.g. applied to the real world environment or evaluated by a teacher, and further refined if the test fails. During the RETAIN phase, useful experience is retained for future reuse, and the case base is updated by including the newly learned case or modifying some of the existing cases. The CBR process could be therefore depicted as follows:

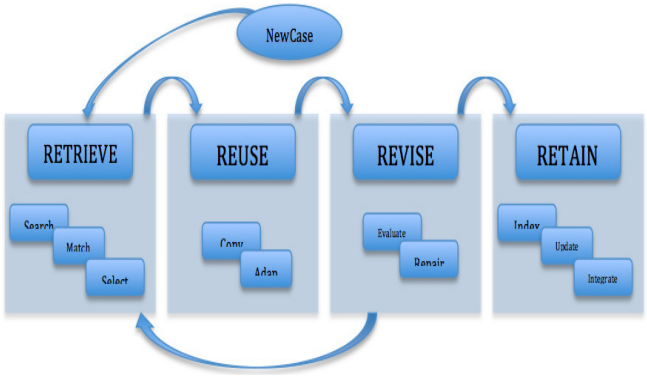


Fig. 2 A CBR model

The application of CBR to support the operation of an ACM system may, for example, proceed as follows:

An event is triggered by e.g. scanning a new document found in incoming mail, initiating a new business case that immediately enters the system. The document is analyzed via an automated OCR process to roughly determine its content, and a new case is opened. Now, the relevant CBR functions may be mapped for processing the case within the ACM system.

RETRIEVE

1. As a first step in the process, the available case data (metadata, content, case type, solution template proposed, etc.) is read. All these data are offered in a legible format by the basic ACM system functions.
2. The next step is to search the case repository in an attempt to find historical cases whose characteristics match the case under examination as closely as possible (metadata, content, document classes, etc.).
3. An initial match: the system selects the cases/information whose characteristics most closely resemble those of the case being examined: similar metadata, business cases, case domains, etc. Unstructured data analysis tools (e.g. asset correlation testing with the use of IBM Watson Content Analytics) can be used to refine the search.

The resulting algorithm reproduces the first functional area (phase) of CBR.

REUSE

1. After selecting a set of similar business cases (based e.g. on such criteria as the sequence of tasks or user activities from the model case retrieved from the case base), it is possible to pick a case processing template created by the

knowledge worker dealing with the previous case and apply the template in handling the new case.

2. In addition, if any documents have already been created in dealing with the case (e.g. replies, clarifications, or notifications conveyed to a customer), these can be included in the file for the case being handled.

These CBR functions can quite naturally become an add-in to, or an extension to, the basic functions of an ACM system. They not only accelerate the decision making process (case processing) but also account for better, more relevant decisions.

REVISE

The case manager software provides analytical tools that can be used to draw specific conclusions from information artifacts related to the case, which may include unstructured and/or structured information.

The third phase of the CBR model involves an evaluation of the extent to which the choice of stored case files (data, information, results) really matches the new case. Within an ACM system, this process can be supported with the use of business analysis tools [7]:

1. Detailed analysis and decision making improvement tools that can be used to optimize case processing in both general and specific terms. An ACM environment supplies analytical tools that help form very specific and detailed conclusions based on case-related information, whether structured or unstructured:

- At the level of individual users, such analyses make it possible to prioritize, assign tasks, and make decisions regarding individual cases.
- At the general level, such tools can help identify certain patterns and trends across a group of cases or evaluate the impact that each of the cases might have on specific organizational units or departments.
- With such information available, managers are able to proactively optimize performance, for example by changing work allocation, hiring additional experts, providing additional information on particular cases, or improving the quality of training.

Although the outcomes of the REVISE phase can bear very positively on the performance of the knowledge workers processing particular business cases in an ACM system, market research data, including ACM usage statistics, show that this phase will not be actually applied in each business case, because it places high demand on the system.

RETAIN

The purpose of this phase of CBR is to update the business case base/repository with information on the history of processing and solving a particular case [3]. Here, ACM will automatically transfer a case into the archive, and store all related documents and metadata in the repository. Once CBR comes into play, the algorithm or workflow path for a particular type of cases is likely to be changed permanently based on previous business cases. When the success statistics are high (accurate decisions, short problem solving times), the ACM system may, with the use of Case Based Reasoning, modify the business rules related to the processing of a particular type of cases. The outcome of the final phase of CBR is that it permits the ACM system to record a history of operations involved in case processing (e.g. task execution, decisions, communication with experts), thus converting restricted-access knowledge into open-access knowledge and expanding the organization's intellectual capital.

A company's body of knowledge is partitioned and distributed among staff and across worker groups; before it can be brought to productive use, it has to be properly organized. Any modern enterprise needs to have a Knowledge Management system, which can be described as a complex blend of understanding and experience, explicit and tacit knowledge, material and social technology [8].

The following are the principal goals of Knowledge Management in organizations:

- to make the most of the knowledge that is already available within the organization, and
- to create new knowledge.

It has taken some time for companies and researchers to realize that, besides data as such and besides information that can be interpreted by humans, there exists another vital resource that becomes increasingly crucial to a company's performance but cannot be captured and managed via standard information management methods – and that is knowledge. It can be either explicit knowledge, readily accessible e.g. from an Internet portal, or tacit knowledge that resides in the staff's minds and originates in their individual experience, training and talent.

VI. APPLICATION EXAMPLES

The Knowledge Worker perspective:

Under modern ACM systems, knowledge workers (KWs) will start work without any templates or ready-to-use solutions –with a blank ACM system alone. If they wish so, they can continue to work in that way forever, adding case by case. In the beginning, each case seems different from all other cases. However, as work becomes repetitive, individual knowledge workers will learn to identify snippets of cases that they could convert to personal templates, and reuse. KWs can thus benefit from CBR methodology that has been tailored to the duties they execute within an ACM environment.

For example, on having handled several similar cases, KWs in the back office will recognize that some software checks appear regularly, so it would be best to include them in a template. Furthermore, it should be borne in mind that their reasoning on future cases is augmented through learning from previous cases.

KWs understand that if they can make a template available to their colleagues, they will be able to ask their co-workers once in a while to perform the checks for them – and save some time in this way. Therefore, they will search for a case that contains such checks, copy the part into a new template, and edit the template to provide instructions that other KWs can follow. KWs share their knowledge by publishing that template across the library section for their group, so that other KWs can access it. If a similar case comes up, KWs can copy the template into their case. This example highlights all of the CBR phases at once. In effect, KWs can save some of their time while at the same time sharing their expert knowledge and giving guidance to the other team members through case patterns/templates.

Users of templates can rate them, tag them, and make suggestions for improvements. A template can be promoted to a policy status in order to gain more visibility to KM. The CBR process ensures that templates are not promoted to policies until they have been reviewed and approved by the participants and parties involved. The same is true of discarding templates/policies that are no longer in use. Hence, none but practically proven cases can become templates, and the set of templates is constantly improved: new templates are created on an as-needed basis while obsolete templates are disposed of. This implies that CBR is adapted in iterative cycles. As a consequence, the template library can be adjusted to new processes and new business situations as necessary. This can be accomplished by combining ACM features with the CBR idea to automate a case processing solution.

The Manager perspective:

A manager or team leader needs a workspace that contains team goals and team sprints/milestones, while the team members' personal goals and milestones remain in their individual workspaces. Most of the manager's work is done in the CBR area of REUSE and REVISE.

Since the knowledge work environment is characterized by frequent changes, managers need to have an analytical method of evaluating case data. Where changes occur very often, it is very important to be aware of how many goals have already been achieved, what percentage of goals have been altered, and where the bottlenecks or areas of high goal volatility are. Managers have to be able to instantly establish the reasons why a given case is not progressing, and find out who is responsible for the holdup. Hence, the ability to mine the goals and cases is of great relevance to managers. The key contribution of analytics is brought by making relevant case data available. CBR, on the other hand, having mapped a previous solution to the target problem, takes care of testing the solution in real world settings (a simulation may be performed) and, if necessary, revising it.

An ACM system has an unquestionable merit in coping with staff turnover. Past shared cases, such as e.g. customer service cases, are readily available for retrieval; at the same time, best practices are available in the form of templates and policies. Understandably, most knowledge workers are not inclined to share their experience and expertise unless it directly benefits and speeds up their own work; a degree of mutual trust is an obvious prerequisite for sharing these. Software technology that does not account for the unpredictability of cases is not fit for the purpose. Workspaces offer the right means to protect data, while at the same time allowing the sharing of all that is needed.

VII. The possibility of extending the application of CBR to ACM

Parts of Case Based Reasoning methodology have been used to fuel techniques for retrieving literal information, delivering performance superior to traditional databases. In this way, two new technologies supporting team work have emerged, i.e. Structured Contextual Search (SCS) and Dynamically Contextualized Knowledge Representation (DCKR).

These days, a number of software vendors promote contextual search and natural language search that is informed by the context of information comprised in knowledge bases. Transcending the search paradigm that relies on keywords and connectors, these techniques create room for users to sophisticate their searches toward a more elaborate and more effective approach. A search is considered "contextual" and "structured" when it meets the following criteria:

1. the context of documents stored in the system is taken into account;
2. it is the context that guides the entry, as well as comparison and selection, of documents [9].

It seems that this approach is widely accepted. IBM incorporates natural language processing and unstructured data analysis components into the core of ACM (IBM Case Manager). IBM Watson is a tool that can read and understand natural language, which is key to analyzing unstructured data and hence an invaluable asset in a world where 80 percent of case data are unstructured. Watson enables users to perform unstructured analysis based on a structural pattern detection process.

ACM software packages contain suites of personalizable tools providing whatever organizational and technical means it takes to raise organizational competence, improve the staff's education and learning capability, and boost collective intelligence. It supports the development and use of state-of-the-art mechanisms for semantic content analysis and industry-specific glossaries aiding communication among knowledge workers within an organization [8]. Owing to enhanced text analysis techniques, it makes it possible to discover trends, patterns and relationships within unstructured data as well as within related structured data. The resulting observations become part of organizational knowledge and can be used in decision making, forecasting, and setting business targets. In ACM environments such as

the IBM Case Manager, the user interface and the system vocabulary are customizable and can be adapted to the language specific to a given professional/business area (e.g. medical or other discipline-specific terminology).

The findings of a survey conducted by the authors indicate that the most frequently used creative problem solving tools include:

– context-sensitive help along with access to historical data and similar cases,

– group work support tools, such as discussion forums or (widely popular) instant messengers.

The integration of IBM Content Analytics Watson with IBM Case Manager enables the crawler to link to the most relevant data. Through repeated use, tracking feedback from its users and learning from both successes and failures, Watson gets increasingly smart over time – which also overlaps with the CBR concept.

VIII. Summary

The paper presents the idea of applying the CBR method in the daily activities of a knowledge worker, thus enhancing the performance of an ACM system. It outlines a theoretical underpinning for the use of Case Based Reasoning to support business decision making in case processing. The concept is illustrated with practical examples and a discussion of design implications.

Hopefully, the paper has demonstrated that the application of Case Based Reasoning within ACM can accelerate access to information that is critical to making reasonable business decisions. The integration of Case Based Reasoning into ACM facilitates the exploration of solutions to business cases, which translates into streamlining the problem solving process and, consequently, into making better and more timely business decisions. This indicates that the concept can be beneficial both from the knowledge worker and the middle management perspective.

The proposed incorporation of CBR into ACM systems can provide substantial additional support to managers and knowledge workers, improving the rationality of the decision making process, reducing the risk of decisions made under uncertainty, hence increasing the chances of success. Based on their preliminary research findings, the authors anticipate that the CBR approach might be useful in supporting strategic decision making, especially under Adaptive Case Management systems.

In addition, by delineating a trajectory for optimizing ACM through the incorporation of AI methods, the paper seeks to initiate discussion of the roadmap for the future evolution of ACM systems.

References

[1] A. Abecker, S. Decker, F. Maurer, *"Organizational memory and knowledge management,"* Information Systems Frontiers 2(3–4), 2000, pp. 251-252.

[2] A. Aamodt, M. Nygård, *"Different roles and mutual dependencies of data, information, and knowledge—an AI perspective on their integration,"* Data and Knowledge Engineering 16, 1995, pp. 191-222.

[3] A. Aamodt, E. Plaza, *"Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches,"* AI Communications, IOS Press, Vol. 7(1), 1994, pp. 39-59.

[4] Ł. Osuszek, *Podejście procesowe jako klucz do optymalizacji pracy nowoczesnych przedsiębiorstw.* PTI, Krajowa Konferencja Inżynierii Oprogramowania, 2013.

[5] K. D. Swenson (Ed.), *Mastering the Unpredictable: How Adaptive Case Management Will Revolutionize the Way that Knowledge Workers Get Things Done,* Chapter 1 "The nature of knowledge work." Tampa: Meghan-Kiffer Press, 2010.

[6] A. Heylighen, H. Neuckermans, *"Case Base of Case-Based Design Tools for Architecture,"* Computer-Aided Design, 2001, pp. 1111-1122.

[7] M. L. Maher, A. Gomez de Silva Garza, *"Developing Cased-Based Reasoning for Structural Design,"* IEEE Expert Intelligent Systems & Their Applications, June 1996, pp. 42-53.

[8] E. Skrzypek, *"Wpływ zarządzania wiedzą na wartość firmy"* in E. Urbańczyk (Ed.), *Zarządzanie wartością przedsiębiorstwa w warunkach globalizacji. Wybrane zagadnienia.* Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, 2001.

[9] H. C. Hoeschl, V. Barcellos, *"Artificial Intelligence and Knowledge Management,"* in M. Bramer (Ed.), Artificial Intelligence in Theory and Practice, Boston: Springer, 2006, pp. 11-19.

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 225–232

DOI: 10.15439/2015F80
ACSIS, Vol. 6

Creating an online art exhibition: The impact of online context on the Internet user's experience and behaviour

Urszula Świerczyńska-Kaczor
The Jan Kochanowski University
ul. Żeromskiego 15, 25-369
Kielce, Poland
Email: swierczynska@ujk.edu.pl

***Abstract*—The paper aims to contribute to the discussion about the implementation of virtual art galleries on the websites of art museums and individual artists. This paper focuses on the analysis of the impact of online context on the Internet user's experience and the user's recommendation of the website. The research problem was discussed with reference to the results of an empirical study. The present empirical study led to the following conclusions: 1. the confirmation of the hypothesis about the positive impact of enhanced context on the respondent's perception of the art presentation; 2. the rejection of the hypothesis about the positive impact of enhanced context on the viewer's recommendation of the virtual gallery; 3. the rejection of the hypothesis about the mediating role of interest in art on the viewer's preferred context of virtual art gallery.**

Introduction

NOWADAYS many museums integrate the online presentation of digitalized art images within their website, offering Internet users the opportunity to visit and search through online virtual art galleries. The online art galleries are also an essential feature of the websites of many individual artists who 'exist' and present their work mostly, or exclusively, online. So far, there are few research studies which compare the art experience in the traditional museum with the online art experience. However, some studies suggest that virtual art galleries may be perceived as unsatisfactory: the artworks in the museum environment were found to be more arousing, positive, interesting and liked, and also better remembered, by viewers compared to the computer presentation (Brieber, Nadal and Leder 2015) and the online visit was perceived as less pleasurable and more passive than the actual visit to a traditional museum (Jarrier & Bourgeon-Renault 2012). Therefore, on one hand there is an obvious need for developing online art galleries for promotional, informational and educational purposes, on the other hand – it seems that the existing online artworks exhibitions do not fully meet the needs of potential viewers.

This paper focuses on the analysis of the impact of online context which influences the Internet user's experience with the online exhibition and user's intention to recommend the website of art gallery. The paper points to the results of an empirical study based on the implementation of microfictions as a context for the online art exhibition.

The paper is organized as follows: in the next section, the research problem was discussed with reference to a literature review. This part of the article explicitly illustrates that the research problem should be approached from different angles, and the research needs to be based on the integrated framework built within (at least) the psychology of art, the neuroaesthetics and managerial studies. The following section presents the procedure of the empirical study, which aimed to 'capture' the connection between the enhanced context of the virtual art gallery and the attitude and behaviour of the respondent. For the purposes of empirical study the virtual art gallery with two options of art presentation (artworks with standard information and the artworks with microfictions) was developed. The article ends with a discussion summarizing the empirical results, and indicating the managerial implications and future research.

The Problem

The art museum/gallery website creates the online context of artworks presentation and therefore frames the artwork in a way which influences the perception of the viewer. For online art galleries the content of art presentation usually includes metadata about the artwork such as: name of the artist, the medium, creation date and the size of the artwork. What if the virtual art gallery embeds more than 'standard' information e.g. music which accompanies the viewing of particular fine art work, a poem or other form of artistic expression? Does it lead to positively enhancing the Internet user's experience? Does this 'mixture' of art objects from different artists - musicians, painters, graphic artists, poets or writers - create positive synergy in the process of customer experience? It is not unusual to enhance the customers' experience by employing multimedia (e.g. sound, music, additional art-works such as photographs or even theatrical performance) in order to enrich the customer's experience during the visit to the physical art museum or gallery.

However, the websites of traditional galleries do not usually incorporate such extraneous elements to enhance virtual tours.

The present study focuses on analysing the impact of the context, which includes microfictions as the 'enhanced frame' for online art galleries. Microfictions (or flash fiction, sudden fiction, minute stories, short-shots) are a contemporary miniature narrative genre of literature, which are (usually) shorter than 700 words (Nelles 2012). Therefore, the main research problem is stated: to what extent does the enhanced context of art galleries impact on the online viewer's experience and behaviour? To investigate this connection, at least three different constructs should be taken into account:

1. The design of the website with the online art exhibition, including graphical and multimedia elements.
2. The outcomes of the viewer's experience:
    a. Outcomes connected with art appreciation;
    b. Outcomes connected with the viewer's perception of gallery/museum brand;
3. Viewer characteristics e.g. interest in art, gender, age, motivation to visit a virtual art gallery/museum

A. *Outcomes of experiencing the online art*

In the literature, the models aiming to explain the viewer's art experience varied as the researchers implemented different frameworks in their analyses. For example, the model first proposed by Leder, Belke, Oeberst & Augustin in 2004, and then further discussed in the paper of Leder & Nadal (2014) provides the integrated framework of the process of understanding the aesthetic appreciation of art, and points to aesthetic judgment and aesthetic emotion as the different outcomes of art experience (Leder & Nadal 2014). The model proposed by Pelowski & Akiba (2011) emphasizes the transformative impact of art, the initial disruption, the subsequent meta-cognitive reflection and conceptual change. A relatively new stream of literature, based on the implementation of neuroimaging techniques (e.g. functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG)), seeks for the explanation of art experience at the neurological level: showing the way in which particular art stimuli activates the human brain (see the interesting research presented by Cela-Conde, Agnati, Huston, Mora, & Nadal 2011).

A review of studies from literature indicates that the researchers include and measure different constructs within the 'general dimension' of viewer's art experience, as for example:

- the construct of aesthetic appreciation, the construct of appraised ability to understand the paintings (Swami 2013);
- the dimensions of liking, the interest and emotional valence (Gartus and Leder 2014);
- the dimension of liking, interest, understanding and ambiguity (Brieber, Nadal, Leder, Rosenberg 2014)

The stream of literature of the psychology of art or neuroaesthetic does not focus on the relation between the viewer's art experience (on traditional or online market) and the impact of this experience on the perception of the art gallery brand which displays the artwork (although the problem – how the image of the gallery influences the viewer's perception of artwork was investigated e.g. Kirk, Skov, Hulme, Christensen, & Zeki (2009)). Within the scope of under-researched topics there are (at least) the following connections:

- the online art experience and the patron's attitude to the brand (e.g. perception that the brand of a particular artist is trustworthy);
- the connection between the online art experience of the viewer and the viewer's behaviour connected with the possible purchase e.g. intention to buy the ticket or artwork;
- the online viewer's intention to recommend the art gallery brand as a result of visiting the art online gallery;

The above mentioned business outcomes as the result of the patron's visit to a virtual gallery are linked with broader constructs such as the art gallery brand equity and the customer equity. So far, there are few research papers in this field of study, and many issues need further investigation, including fundamental questions about the elements which constitute the brand equity of a virtual art exhibition, and about the drivers of customer equity for the website with a virtual art exhibition.

With regards to the traditional market, Camarero, Garrido, & Vicente (2010) indicate the loyalty, perceived quality, brand image and brand values as the factors influencing the brand equity for art exhibitions. However, it seems possible that the patrons may perceive the website and the traditional brand of art museum/gallery differently. For example, on the basis of direct visits and seeing a particular exhibition, the patron perceives the brand of the art museum negatively (offering too few exhibition items, low quality of organizing the service etc.), however the website with a virtual art gallery of the same art museum is perceived as very well-organized and well-developed. This example illustrates that the viewer's perception of virtual gallery brand and the 'traditional' art gallery brand may be partly independent constructs, and the relation between these constructs may be mediated by other factors, for example the viewer's previous experience.

In the present empirical study the viewer assessed the virtual gallery – with or without the microfictions– in the following dimensions:

- the dimensions connected with art experience such as of 'liking the presentation' and 'the richer experience'
- the dimension connected with the business outcome – 'the intention to recommend the virtual gallery'.

B. The context of online art presentation

Contrary to the formalist views of art, nowadays the context in which the artwork is viewed has been recognized as a significant factor influencing the viewer's experience with art. Its impact on art perception are well-illustrated by, for example the Duchamp's 'readymade' artworks ('Bottle Rack', 'Fountain', 'Bicycle Wheel') or the Pop Art movement (see e.g. discussion in Leder & Nadal 2014, Gartus & Leder 2014).

The context in which artwork is presented has been investigated and discussed in the literature, indicating the impact of different contextual aspects on the viewer's experience. The study of Leder, Carbon, Ripsas (2006) pointed to the role of elaborative and descriptive titles of abstract and representative paintings**.** Swami (2013) examined the effects of different types of information (elaborate, content specific information, broad genre information, titular information, no contextualizing information) and different art styles (abstract artworks, representational paintings) on understanding and aesthetic appreciation. Kirk, Skov, Hulme, Christensen, & Zeki (2009), using the neuroimaging technique, investigated the perception of the image labelled either as an artwork belonging to the art gallery or an image generated by a computer. In another study with usage of fMRI, Cupchik, Vartanian, Crawley, & Mikulis (2009) investigated the differences in perception of art images under the pragmatic and aesthetic conditions.

Despite the difference between the presentation of an art object in a physical museum and the laboratory environment, most of the studies with reference to art experience were conducted within laboratory settings. Therefore, it seems plausible, that the results obtained in the laboratory settings may be relevant to the discussion about the creation of online art exhibitions.

In the online environment, the context of presentation is created not only by information about the art, but also by the scope of the elements constituting 'the design' of the website. The following elements of a museum website may influence the perception of the Internet user: 1. Content, 2. Presentation-Media-Format-Appearance, 3. Usability, 4. Interactivity & Feedback, 5. E-Services, 6. Technical (see the framework for evaluating the museums' websites developed by Pallas and Economides 2008). Among these elements, the option of creating personal digital collections is particularly important as it serves: for example, to facilitate the visit in the museum in person or to facilitate learning (Marty 2011).

The important distinction between the museum websites lies in the format presentation based on 2D websites versus the 3D virtual environment. Nowadays the object of art may be 3D digitalized by the implementation of different techniques (for example laser scanning techniques; shape from structured light, from silhouette, from video, from shading, and shape from texture – Pavlidis, Koutsoudis, Arnaoutoglou, Tsioukas, & Chamzas 2007), and the Augmented Representation of Cultural Objects (ARCO) system allows the creation of virtual 3D museum exhibitions (Sylaiou, Mania, Karoulis & White 2010). Some individual artists use the 3D virtual world's environment, such as Second Life, to create a virtual gallery which allows the viewer to be immersed in the space. In the case of fine art most digitalized exhibitions are based on the traditional 2D presentation on the website. The present empirical study also focuses on the 2D presentation format, and the context created by the implementation of additional element – microfiction.

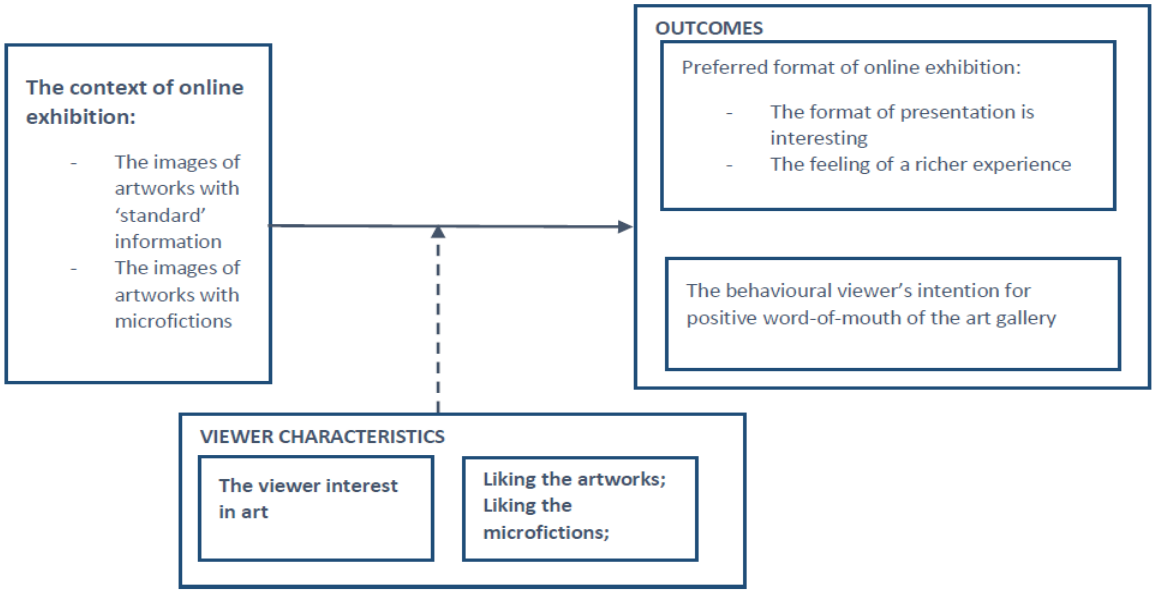


Fig. 1 The framework for empirical study

C. The viewer characteristics

The one of the crucial factors connected with viewer characteristics is the viewer's motivation. The museum websites aim to fulfil the different Internet user's informational needs: obtaining information relevant to the process of planning the visit to the physical museum, self-motivated research for specific content information, looking for the information connected with assigned research or being engaged in casual browsing (Skov & Ingwesen 2014). In the case of the individual artist (painter, graphic artist, photographer) while the website often serves as the sales channel, the motivation of potential customers may also be connected with a purchase.

The viewer's characteristics which may have an impact of the viewer's art experience are, for example aesthetic fluency (Swami 2013), current affective state and specific art interest (Gartus, Leder 2014), and expertise in art (Leder, Nadal 2014). In the case of online galleries the spectrum of factors should be widened to incorporate aspects related with Internet and computer usage. For example, Corredor's (2006) study with reference to museum websites points out that the general previous knowledge (in the field corresponding with museum) influences the goal setting process and the usage of the content of the website. For the virtual gallery the viewer's 'general' level of computer literacy may be an important factor, especially in the situation where the lack of computer skills may hamper the viewer's experience.

THE EMPIRICAL STUDY: HYPOTHESES, THE PROCEDURE AND THE RESULTS

A. Hypotheses

The present study aims to investigate the connections between the context of presentation – the artworks with or without the microfictions - and the following outcomes (Fig. 1):

- the viewer's perception of the virtual gallery with particular context as an interesting way of presentation;
- the viewer's feeling of 'rich' experience connected with viewing artwork in the particular context;
- the viewer's intention to recommend the virtual art gallery to a friend;

The other factors included in the present study are: the interest in art, and the level of liking the artworks and microfictions.

In the present study, the following hypotheses were tested:

H1: The online enhanced context with microfiction leads to a more positive attitude of the respondent towards the website in the dimensions of liking the website and offering a 'richer' experience.

H2: The online enhanced context with microfiction leads to the behavioural intention of the viewer: the intention to recommend the website to a friend

H3: The viewer's interest in art is a factor which influences the preference of the virtual gallery by the viewer.

B. Participants

The study was conducted in Poland, in January 2015, among 50 young respondents (32 women and 18 men), age 20-35 (the group of respondents was homogenous as to the age factor). The respondents were management students, without formal background knowledge of art.

C. The procedure

The website of a virtual galley was developed for the purpose of the present study. In the introduction to the study the respondent was informed that the website was a part of a future virtual tour being developed by a new online art gallery. On the introductory page the respondent was asked to visit the next two pages, titled 'Virtual Gallery' and 'Virtual Gallery with Microfictions', and then to fill in the online questionnaire. Five artworks of a local artist were selected and the images were presented on the website. On the page 'Virtual Gallery' the images of artworks had a 'standard' description: the artistic pseudonym of the author of the artwork, the title of artwork, the medium of which the artwork was created and the size of the original work (length and height in centimetres). The technique of the artworks varied: acrylic on canvas (one artwork), pastels (two artworks) and graphics (two artworks). On the page titled 'Virtual Gallery with Microfiction' each image of artwork was accompanied by a short microfiction. The topic of each microfiction aligned with the main concept of the artwork. The length of the microfictions was between 22 and 84 words, the name of the

TABLE I.
THE RESPONDENTS' PREFERENCES FOR THE CONTEXT OF ARTWORKS PRESENTATION; STATISTICALLY SIGNIFICANT DIFFERENCE MARKED *

	All respondents N=50			**The group of respondents – similar liking of artworks and microfiction n = 44**		
	standard presentation	**artworks with microfictions**	**I don't know**	**standard presentation**	**artworks with microfictions**	**I don't know**
The more interesting presentation of artworks is on the website with…	20%*	70%*	20%	16%	73%	11%
My experience connected with visiting virtual art gallery was richer while I visit the website with …	24%*	64%*	24%	18%	68%	14%
I would be more likely to recommend the visit to the museum website to my friends if the museum website included the artwork presentations as at the page of the website with …	24%	54%	24%	20%	55%	25%

author of each microfiction was clearly indicated on the website.

The procedure of the study was followed as:

a) First step: the respondents were introduced to the concept of the study.
b) Second step: The respondent saw the page 'Virtual Gallery'
c) Third step: The respondent saw the page 'Virtual Gallery with Microfiction'
d) Fourth step: Respondent filled in the questionnaire. The online questionnaire was divided into three groups of questions.
    - First group of questions were directly related to the virtual gallery.
    - In the second group of questions, respondents were asked to visit the digital exhibitions of the National Museum in Warsaw. After visiting the website, the respondents filled in the part of the questionnaire which refers to the respondents' attitude to the process of creating virtual galleries by museums.
    - The third group of questions refers to the profile of respondents. Apart from gathering data about their gender and age, the questions refer to their interest in 'general' art and their interest in visual art.

D. Results

As the respondent's attitude to the art presentation may be influenced by liking the artworks or microfictions themselves, the respondent was asked to rank the level on which she or he generally likes the artworks or microfiction: as the general feeling about 'liking the microfiction' or 'liking the artworks'. There was no significant difference in respondents' feeling of liking of artwork and the microfictions: about half of the respondents (46%) liked the presented artworks and a similar percentage of respondents (50%) liked the microfictions. The averages of 'liking' were 3.34 and 3.42 for the artworks and for the microfictions respectively.

The majority of respondents – 88% - ranked their liking for the artworks and microfictions similarly – the difference between 'liking' wasn't greater than one point in the scale of 1 to 5 points. The group of 6 respondents (12% of the sample) preferred one of the presented forms: the difference between the 'liking' was greater than 1 point.

The interest in art and in visual art varied among the respondents:

- The 38% of respondents declared that they are not interested in art, with reference to the visual art – also it was the group of 38% respondents
- The 26% of respondents declared their interest in art, with reference to visual art – 30%
- The group of respondents rated their interest as 'neutral' (3 points on the 1-5 points scale) in art was – 36%, in visual art – 32%

The study led to the following results (see Table I, Table II, and Table III):

1. The respondents perceived the artwork presentation with microfictions as much more interesting than the 'standard' presentation. The webpage with microfiction was selected as favourable in this dimension by 70% of respondents compared to the 20% of the respondents preferring the 'standard presentation' (the statistically significant difference $p<0.05$).
2. 64% of respondents stated that the virtual gallery with microfiction 'offers a richer experience' compared to the 24% of respondents who preferred the 'standard' presentation (the difference is statistically significant $p<0.05$)
3. If the museum website included artworks and microfiction, the 54% of respondents would recommend their friends to visit this museum website, compared to the group of 24% who would recommend the website with the standard presentation (no statistically significant difference)
4. Relations between the variables 'more interesting presentation', 'offering richer experiences' and 'the intention to recommend' were analysed within two options: the respondent who chose the website of microfiction or the respondent who chose the other answers: gallery without microfiction or 'I don't know':
    a. The variable of 'more interesting presentation' and the variable 'offering richer experiences' are strongly positively

TABLE II.
THE CORRELATION BETWEEN THE VARIABLES [OPTIONS: 'THE WEBSITE WITH MICROFICTION' OR THE OTHER ANSWERS: {THE WEBSITE WITHOUT MICROFICTION OR 'I DON'T KNOW'}]

'the most interesting presentation' – 'richer experience'	Positive, statistically significant, test chi-square $p<0.05$, Fi=0.60, Tetrachoric correlations=0.82
'the most interesting presentation' – 'recommendation to friend'	Positive, statistically significant - the chi-squared test $p.<0.05$, and Fi=0.62, Tetrachoric correlations=0.88
'the richer experience' – 'recommendation to friend'	Positive, statistically significant - the chi-squared test $p.<0.05$, Fi = 0.39, Tetrachoric correlations=0.59
The group which is not interested in art (1, 2, or 3) and the group which is interested in art (4-5)	There is no difference between the groups in the dimensions: 'liking virtual galleries', 'richer experience' or 'recommendation to friend'

correlated (the chi-squared test p.<0.05, Fi=0.60, Tetrachoric correlations=0.82).

b. The variable 'more interesting presentation' strongly positively correlates with the variable 'intention to recommend' (the chi-squared test p.<0.05, and Fi=0.62, Tetrachoric correlations=0.88)

c. The variable 'much richer experiences' and the variable 'recommendation' are weakly positively correlated (the chi-squared test p<0.05, Fi=0.39, Tetrachoric correlations=0.59)

d. There is no difference between the assessment of the online exhibition with microfiction or without between the group of respondents who liked the artworks (first group: 4 or 5) and those who disliked the artworks (another group – 1,2 or 3 points) (chi-square test, p>0.05)

e. There is no difference between the assessment of online exhibitions between the group of respondents who are interested (4-5) and those who are not interested in art/visual art (1, 2 or 3).

5. The results with reference to the questions about the website of the National Museum of Warsaw point out that:

a. The majority of respondents (62%) agreed with the statement: "the online museum presentation and digitalization of museum objects allow me to look at the museum exhibition in a way which suits me";

b. The majority of respondents (68%) supported the engagement of museums in creating online galleries;

c. The online museum did not necessary encourage the respondents to visit the traditional museum;

To sum up:

- The hypothesis H1: The online enhanced context with microfiction leads to a more positive attitude of the respondent towards the website in the dimensions of liking the website and offering a 'richer' experience – was confirmed.
- The hypothesis H2: The online enhanced context with microfiction leads to the behavioural intention of viewer: the intention to recommend the website to a friend – was rejected
- The hypothesis H3: The viewer's interest in art is a factor which influences the preference of the virtual gallery by viewer – was rejected

E. Limitations

The following points underline the limitations of the present study:

1- Although the procedure of the study allows respondents to express their opinion about the preferred format of presentation, the present study is not based on the experiment aiming to make comparisons between the group of respondents who viewed the 'virtual gallery' with standard data and a different group of respondents who viewed the 'virtual gallery with microfiction'.

2- The stimuli: the choice of artworks and the microfictions was arbitrary, and the style and content of presented artworks varied (graphics, pastels, and acrylic). The procedure did not include the respondents' evaluation on what level the artworks and microfictions corresponded within an artistic theme.

3- The results refer to the 'general' website presentation, not to particular artworks. The respondents did not evaluate each artwork and microfiction.

4- Apart from the demographic data, the viewer's characteristics included few other variables.

5- The sample: non-random sample, the sample of respondents was relatively small, and homogenous in age.

DISCUSSION

The results of this present study should be framed within much broader topics: How to enhance the Internet user's experience during the visit to the website with virtual art exhibition? Does this enhancement of the user's art experience correlate with strengthening the relationship between the Internet user (possible customer) and the museum/gallery brand? Does the virtual gallery influence the brand equity and customer equity for art exhibitions? This scope of research questions should also be formulated and addressed with reference to the groups of sponsors and donors, as these groups play a crucial role in the existence and development of cultural art organizations.

TABLE III.
THE RESPONDENTS' ATTITUDE TO THE DIGITAL MUSEUMS; STATISTICALLY SIGNIFICANT DIFFERENCE MARKED *

	All respondents N=50		
	Negative (1-2)	**Neutral (3)**	**Positive (4-5)**
In my opinion, the online museum presentation and digitalization of museum objects allows me to look at the museum exhibition in a way which suits me	12%*	26%	62%*
The online museum encourages me to visit the traditional museum	16%	36%	48%
Museums and galleries should aim to create online exhibitions	10%*	22%	68%*

The present study indicates that respondents assessed the virtual gallery with microfictions as the more interesting form of presentation and offering a 'richer' experience compared to the standard presentation. However, the 'enhanced presentation' does not necessarily lead to recommendations of the online gallery to friends. The present study also points out that the respondents had a positive attitude to the process of digitalization of exhibitions by museums, but this positive attitude was not necessarily connected with the intention to visit museum in person. There was no difference between the choices of the preferred online art gallery between the group of respondents who are and who are not interested in art.

As regards to management implications, the present study indicates that the strategy of enhancing the Internet user's art experience may be based on linking together two different artworks: visual art and literature. Considering the amount of digitalized objects in museums' virtual galleries, it would not be possible to create 'the literature' story for each artwork. However, the museum may select a few of the artistic objects, and create the 'enhanced literature presentation' as promotional materials. For individual artists the present study also suggests 'telling the story' about artworks as a tool for enhancing the Internet user's experience. Although, this strategy raises the question - to what extent can the museum modify the context without the artist's agreement, if the context may change the meaning of artwork?

In the broader context of brand equity and customer equity, the results of the present study indicate the link between the enhanced context and both constructs. The 'virtual gallery with microfiction' led to the more positive attitude of the respondents, subsequently creating 'better associations' with artistic brand, therefore enhancing the brand equity (see model of brand equity proposed by Aaker (1996)). Moreover, the brand equity, together with value equity and relationship equity, influences the customer equity on business markets (Rust, Lemon, & Zeithaml 2001), and this connection may be similar on the online art galleries market. It is worth noting that another aspect tested in the empirical study – the positive referrals – also may be linked to brand equity (due to loyalty) and to the customer equity as:

- the positive referrals create the indirect value of the customer (see – Ryals 2008)
- the positive referrals may indicate customer's loyalty which influences the customer equity (see – Kossecki 2009)

As it was mentioned above, there are still few research papers investigating the fundamental questions about the elements which constitute the brand equity of a virtual art exhibition, and about the drivers of customer equity for the website with a virtual art exhibition. Therefore future research lies in:

• the field of the more extensive study about the connection between the customer equity and brand equity with reference to the websites of art museum/individual artists;

• how other strategies of modifying the context – e.g. implementing music, sound, poems – influence the Internet users' perception of virtual galleries;

• the relation between the purpose of Internet user's visit and the perception of virtual art galleries. In the present study, the respondents focused on evaluating the online presentation due to the 'artistic' merits. The Internet user's experience may be different if they had, for example educational motivations.

• the dynamic changes and interrelations between the art experience and the patron's visit to the virtual gallery. The process of the art experience can change over the time: the first visit may differ from subsequent visits, therefore the scope of the research questions indicated above should take into account the factor of 'time' as an important variable.

REFERENCES

[1] Aaker D. A. (1996), "Measuring Brand Equity Across Products and Markets", *California Management Review*. Spring96, Vol. 38, No. 3, 102-120.

[2] Brieber D., Nadal M., Leder H. (2015), "In the white cube: museum context enhances the valuation and memory of art", *Acta Psychologica*, 154 (2015), 36-42, doi:10.1016/j.actpsy.2014.11.004

[3] Brieber D., Nadal M., Leder H., Rosenberg R. (2014), "Art in Time and Space: Context Modulates the Relation between Art Experience and Viewing Time", *PLOS ONE*, June 2014, Volume 9, Issue 6, e99019, 1-8.

[4] Camarero C., Garrido M. J., Vicente E., (2010), "Components of art exhibition brand equity for internal and external visitors", *Tourism Management*, 31 (2010), 495-504, doi: 10.1016/j.tourman.2009.05.011

[5] Cela-Conde C. J., Agnati L., Huston J. P., Mora F., Nadal M. (2011), "The neural foundations of aesthetic appreciation", *Progress in Neurobiology* 94(2011) 39–48, doi:10.1016/j.pneurobio.2011.03.003

[6] Corredor J. (2006), "General and domain-specific influence of prior knowledge on setting of goals and content use in museum websites", *Computers & Education*, 47 (2006), 207-221, doi: 10.1016/j.compedu.2004.10.010

[7] Cupchik G. C., Vartanian O., Crawley A., Mikulis D. J. (2009), "Viewing artworks: Contributions of cognitive control and perceptual facilitation to aesthetic experience", *Brain and Cognition*, 70 (2009), 84–91, doi:10.1016/j.bandc.2009.01.003

[8] Gartus A., Leder H. (2014), "The White Cube of the Museum Versus the Gray Cube of the Street: The Role of Context in Aesthetic Evaluation", *Psychology of Aesthetics, Creativity, and the Arts*, 2014, Vol. 8, No. 3, 311-320, http://dx.doi.org/10.1037/a0036847

[9] Jarrier E., Bourgeon-Renault D. (2012), "Impact of Mediation Devices on the Museum Visit Experience and on Visitors' Behavioural Intentions", *International Journal Of Arts Management*, Fall 2012, Volume 15, Number 1, 18-29.

[10] Marty P. F. (2011), "My lost museum: User expectations and motivations for creating personal digital collections on museum websites", *Library & Information Science Research*, 33 (2011), 211–219, doi:10.1016/j.lisr.2010.11.003

[11] Kirk U., Skov M., Hulme O., Christensen M. S., Zeki S. (2009), "Modulation of aesthetic value by semantic context: An fMRI study", *NeuroImage,* 44 (2009), 1125–1132, doi:10.1016/j.neuroimage.2008.10.009

[12] Kossecki P. (2009), Valuation and Value Creation of Internet Companies - Social Network Services (September 26, 2009). Available at SSRN: http://ssrn.com/abstract=1478713 or http://dx.doi.org/10.2139/ssrn.1478713

[13] Leder H., Carbon C.-C., Ripsas A.-L. (2006), "Entitling art: Influence of title information on understanding and appreciation of paintings", *Acta Psychologica*, 121 (2006), 176–198, doi:10.1016/j.actpsy.2005.08.005

[14] Leder H., Nadal M. (2014), "Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode – Developments an challenges in empirical aesthetics", *British Journal of Psychology*, 2014, 105, 443-464, doi:10.1111/bjop.12084

[15] Nelles W. (2012), "Microfiction: What Makes a Very Short Story Very Short?", *NARRATIVE,* Vol 20, No. 1, January 2012, 87-104.
[16] Pallas J., Economides A. A. (2008). "Evaluation of art museums' web sites worldwide", *Information Services & Use,* 28 (2008), 45-57. doi:10.3233/ISU-2008-0554
[17] Pavlidis G., Koutsoudis A., Arnaoutoglou F., Tsioukas V., Chamzas C. (2007), "Methods for 3D digitization of Cultural Heritage", *Journal of Cultural Heritage,* January 2007, 8 (2007), 93-98, doi:10.1016/j.culher.2006.10.007
[18] Pelowski M., Akiba F. (2011), "A model of art perception, evaluation and emotion in transformative aesthetic experience", New Ideas in Psychology, 29 (2011), 80–97, doi:10.1016/j.newideapsych.2010. 04.001
[19] Rust R. T., Lemon K. N., Zeithaml V. A. (2001), "Where Should the Next Marketing Dollar Go?", Marketing Management, 10(3), 24-28.
[20] Ryals L. (2008), "Determining the indirect value of a customer", Journal of Marketing Management, 24(7/8), 847-864.
[21] Skov M., Ingwersen P. (2014), "Museum Web search behavior of special interest visitors", Library & Information Science Research, 36 (2014), 91-98, http://dx.doi.org/10.1016/j.lisr.2013.11.004
[22] Swami V. (2013), "Context Matters: Investigating the Impact of Contextual Information on Aesthetic Appreciation of Paintings by Max Ernst and Pablo Picasso", Psychology of Aesthetics, Creativity, and the Arts, 2013, Vol. 7, No. 3., 285-295, doi: 10.1037/a0030965
[23] Sylaiou S., Mania K., Karoulis A., White, M. (2010), "Exploring the relationship between presence and enjoyment in a virtual museum", International Journal of Human-Computer Studies, 68 (2010), 243-253. doi:10.1016/j.ijhcs.2009.11.002

Joint Agent-oriented Workshops in Synergy

JOINT Agent-oriented Workshops in Synergy is a coalition of agent-oriented workshops that come together to build upon synergies of interests and aim at bringing together researchers from the agent community for lively discussions and exchange of ideas. For the first time JAWS was organized during the 2011 FedCSIS Conference. Workshops that constitute JAWS in 2015 are:

- ABC:MI'15 - 10th Workshop on Agent Based Computing: from Model to Implementation
- MAS&S'15 - 9th International Workshop on Multi-Agent Systems and Simulations
- SEN-MAS'15 - 4th International Workshop on Smart Energy Networks & Multi-Agent Systems

10th Workshop on Agent Based Computing: from Model to Implementation

THE FIELD of agent technology is rapidly maturing. One of key factors that influence this process is the gathered body of knowledge that allows in-depth reflection on the very nature of designing and implementing agent systems. As a result, there is now significant knowledge on how to design and implement them. There is also a deeper understanding of the most important issues to be addressed in the process. Therefore, on the top-most level a progress in development of methodologies for design of agent-based systems can be seen. Furthermore, these methodologies are usually supported by tools that allow not only top level conceptualization but guide the process towards implementation (e.g. by generating at least some code). Next, it can seen that new languages for agent based systems are created, e.g. AML or API Calculus. Separately, tools/platforms/environments that can be used for design and implementation of agent systems have been through a number of releases, eliminating problems and adding new, important features. Resulting products are becoming truly robust and flexible. Furthermore, open source products (e.g. JADE) are surrounded by user communities, which often generate powerful add-on components, further increasing value of existing solutions.

Topics

The Workshop primarily focuses on all aspects of the process that leads from the model of the problem domain to the actual agent-based solution. These aspects will cover both principled approaches and established practices of software engineering aimed at producing high quality software. In this context, research into the application of agent-based solutions to key challenges faced by software engineering (e.g. reduction of costs and delivery times, coping with a larger diversity of problems) will be of primary importance. ABC:MI Workshop welcomes submissions of original papers concerning all aspects of software agent engineering.

Topics include but are not limited to:

- Methodologies for design of agent systems
- Multi-agent systems product lines
- Modeling agent systems
- Agent architectures
- Agent-based simulations
- Simulating and verifying agent systems
- Agent benchmarking and performance measurement
- Agent communication, coordination and cooperation
- Agent languages
- Agent learning and planning
- Agent mobility
- Agent modeling, calculi, and logic
- Agent security
- Agents and Service Oriented Computing
- Agents in the Semantic Web
- Applications and Experiences

Event Chairs

Badica, Costin, University of Craiova, Romania

Ganzha, Maria, University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland

Paprzycki, Marcin, Systems Research Institute Polish Academy of Sciences, Poland

Rahimi, Shahram, Southern Illinois University, United States

Program Committee

Agotnes, Thomas, University of Bergen, Norway

Ambroszkiewicz, Stanislaw, Institute of Computer Science, Polish Academy of Sciences, Poland

Balke, Tina, University of Surrey, United Kingdom

Barseghyan, Artak, Yerevan, Armenia

Botía, Juan, Universidad de Murcia, Spain

Braubach, Lars, University of Hamburg, Germany

Budimac, Zoran, Faculty of Sciences, Univ. of Novi Sad, Serbia

Byrski, Aleksander, AGH University of Science and Technology, Poland

Cabri, Giacomo, University of Modena and Reggio Emilia, Italy

Cervenka, Radovan, Whitestein Technologies AG, Slovakia

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

Fernández, Alberto, Universidad Rey Juan Carlos, Spain

Florea, Adina, University POLITEHNICA of Bucharest, Romania

Gams, Matjaz, Jozef Stefan Institute, Slovenia

Gomez Sanz, Jorge, Universidad Complutense de Madrid, Spain

Goncalves, Ricardo, Uninova, Portugal

Hinchey, Mike, Lero-the Irish Software Engineering Research Centre, Ireland

Ivanović, Mirjana, University of Novi Sad, Serbia

Jamroga, Wojtek, University of Luxembourg, Luxembourg

Jedrzejowicz, Piotr, Gdynia Maritime University, Poland

Jezic, Gordan, University of Zagreb, Croatia

Kaleta, Mariusz, Warsaw University of Technology, Poland

Khorasani, Elham, University of Illinois at Springfield, United States

Koukam, Abder, IRTES-SeT Université de Technologie de Belfort Montbéliard, France

Kruczkiewicz, Zofia, Wrocław University of Technology, Poland

Kusek, Mario, University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia

Leszczyna, Rafal, Gdansk University of Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 237–246
DOI: 10.15439/2015F252
ACSIS, Vol. 6

A Unified Distributed Computing Framework with Mobile Multi-Agent Systems and Virtual Machines for Large-Scale Applications: From the Internet-of-Things to Sensor Clouds

Stefan Bosse
University of Bremen, Dept. of Mathematics & Computer Science,
Robert Hooke Str. 5, 28359 Bremen, Germany

***Abstract*— A novel and unified design approach for reliable distributed and parallel data processing in wide-area and large-scale networks consisting of high- and of low-resource nodes (ranging from generic computers to microchips) using mobile agents is introduced. The development of sensor clouds of the future integrated in daily use computing environments and the Internet is enabled. Agents can migrate between different hardware and software platforms by migrating the program code of the agent, embedding the state and the data of an agent, too. Agent mobility crossing different execution platforms, agent interaction by using tuple-space databases, and agent code reconfiguration enable the design of reliable distributed sensor and information processing networks. The Agent Processing Platform exists in hardware (microchip level), software (embedded system), and simulation. This works adds a JavaScript implementation including client-side browser applications. All implementations are compatibility on operational and communication level. A graph-linked multi-broker service and a distributed co-ordination layer are established for this platform class to provide service ports and the access of the agent platform from the outside in browser applications, which can usually only act as clients and are usually hidden by a private network and firewalls.**

I. Introduction

TRENDS recently emerging in engineering and micro-system applications such as the development of sensorial materials **[3][11]** show a growing demand for distributed autonomous sensor networks of miniaturized low-power smart sensors embedded in technical structures. Multi-agent systems (MAS) can be used for a decentralized and self-organizing approach of data processing in a distributed system like a resource-constrained sensor network (discussed in **[11]** and **[12]**), enabling smart and adaptive distributed information extraction, e.g., based on pattern recognition (e.g., referring **[13]** and **[14]**), by decomposing complex tasks in simpler cooperative agents. It can be shown that MAS-based data processing approaches are scalable from generic computer to single microchip level platforms which can aid the material-integration of Structure and System Monitoring applications. On one hand there are currently only few proposed agent processing platforms that can be scaled to microchip level, and on the other hand there are no unified solutions to integrate these low-resource nodes in large-scale networks and the Internet.

In **[11]** the agent-based architecture considers sensors as devices used by an upper layer of controller agents. Agents are organized according to roles related to the different aspects to integrate, mainly sensor management, communication and data processing. This organization isolates largely and decouples the data management from changing networks, while encouraging reuse of solutions.

The deployment of agents can overcome interface barriers and closes the gap arising between platforms and environments differing considerably in computational and communication capabilities, enabling, e.g., the integration of sensor networks in large-scale WWW applications and providing Internet connectivity, shown in Fig. **1**. This is addressed in this work by using a unified reactive agent-based programming and interaction model, independent of the underlying processing platform. For the proposed advanced agent processing platform architecture there exist suitable hardware (microchip), software (*C*, *OCaML*, *JavaScript*), and simulation model implementations, which can be functionally interconnected in networks creating one big machine. They are compatible on the operational and execution level, thus, agents can migrate between these different implementation platforms.

Agent mobility crossing different execution platforms, agent interaction by using tuple-space databases, and global signal propagation aid solving data distribution and synchronization issues in the design of distributed wide-area networks.

Usually sensor processing and information computation require known world models including mechanical models, e.g., in load monitoring use cases of technical structures. Self-organizing MAS **[2][14]** are useful in unreliable and partially unknown environments, which can overcome world environment and model limitations successfully. Adaptation of the agent behaviour, i.e., based on learning, offers a reliable reaction mechanism in the presence of environmental changes, e.g., changes in network connectivity or node failures, ensuring the QoS. This adaptivity is addressed in this work by a behavioural reconfiguration at run-time, which bases on Dynamic Activity-Transitions Graphs (DATG). Mobility - the ability to migrate an agent processing unit to a different execution platform or node - and autonomy together with a high degree of independency from the processing platform ensure robust data processing in large-scale networks.

It can be shown that agent-based computing can be used to partition complex computations in off-line and on-line parts resulting in an increased overall system efficiency (performance and energy demands), e.g., for Load and Structural Health Monitoring (LM/SHM) systems, outlined in **[2]**.

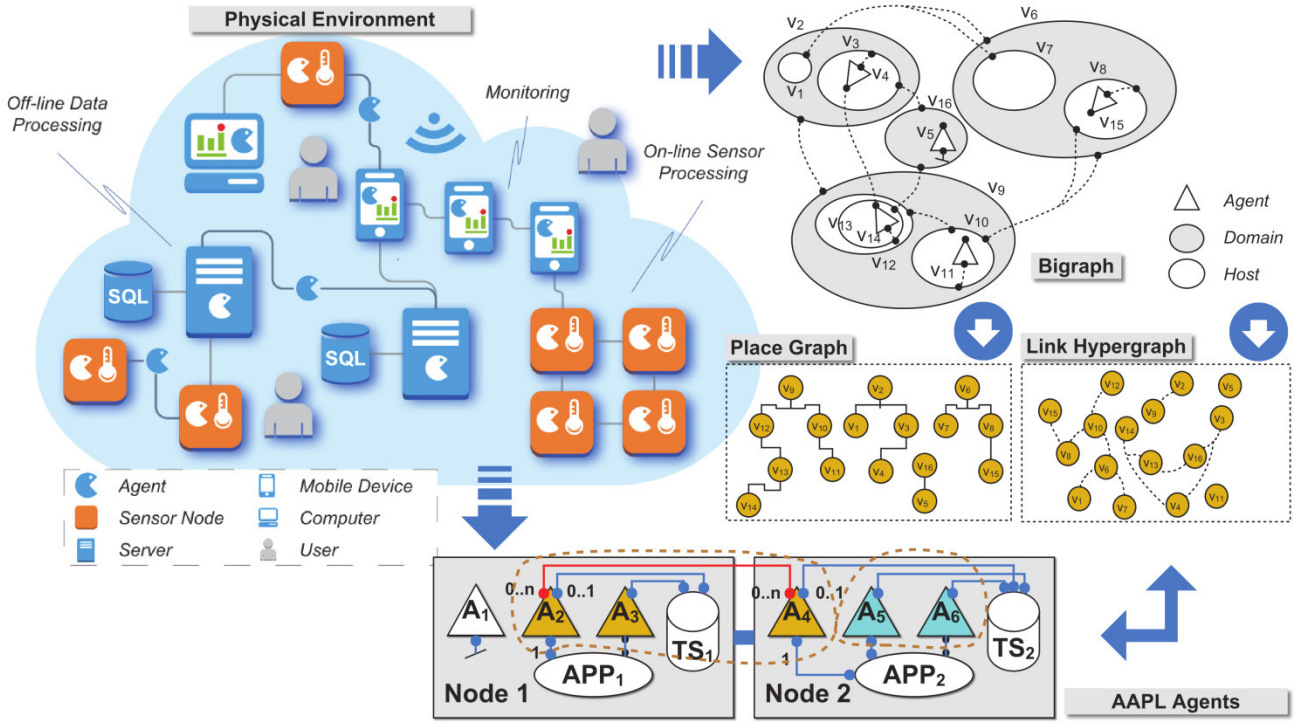


Fig. 1 (Left) Deployment of Agents in Sensor Clouds and Internet Applications (Right) Bigraph, composed of Link and Structure place graphs used for a unified modelling of network environments and networks of networks (Bottom) AAPL agents in the Bigraph Model with a bottom port for the APP link and top port for tuple space and signal link ports. Shown are two connected nodes. [A: Agent, APP: Agent Processing Platform, TS: Tuple Space].

One major goal of the deployment of MAS is overcoming heterogeneous platform and network barriers arising in large scale hierarchical and nested network structures (i.e., networks of networks), consisting and connecting, e.g., the Internet, sensor networks, body networks, production and manufacturing Cyber-Physical System (CPS) networks, shown in Fig. **1** on the left. The large diversity of execution platforms, network topologies, services provided by network nodes, and the programming environments require a unified and abstract behavioural and structural representation model. The Bigraphical model proposed by Robin Milner models the entire "computing" environment with place and link graphs, composing finally bigraphs **[15]**, shown on the right of Fig. **1**. They include agents, and they are offering a unified model and platform for ubiquitous systems and the foundation for an Ubiquitous Abstract Machine, and supporting reconfigurable spaces (dynamic topologies). Bigraphs virtualize communicating processes (agents) and information objects (tuple-spaces), and they originate in process calculi for concurrent systems, especially the pi-calculus **[16]** and the calculus of mobile ambients **[17]** for modelling spatial configurations of networks with a dynamic topology.

The environment consists of places where computation occurs, e.g., computers (processing agents), agents, rooms, buildings, machines, and so on. The links are abstract, providing the possibility of interaction between different places, i.e., transferring of agents and their mobile processes. Agents are treated as active computational units. Places introduce spatial and logical bindings. Bigraphs allow the nesting of nodes and places, natural for many real-world computing environments, and they can be applied for wide reactive systems. All nodes have a fixed number of ports, providing an endpoint for links. Agents have two ports: a processing port link and an interaction (communication) link. Bigraphs, which represents the system state, can be modified by the application of reaction rules, which changes the linking and place relations. Bigraphs can be composed of other bigraphs matching inner and outer interfaces.

A link is a hyperedge connection that connects nodes, outer, and inner names, where names are open linkings that support additional connectivity, i.e., used for the dynamic composition of bigraphs at "run-time". Connectivity not only provides the platform for agent migration between different places, it provides information exchange, which is provided here by place-bounded tuple-spaces and signals. Migration of mobile processes is just another form of interaction with and the modification of the environment.

To adapt this Bigraphical Reactive System (*BRS*) model to a MAS it is necessary to distinguish subjects (entities which can perform actions, the agents) and objects (here data, tuples, tuple-spaces, signals, and processing platforms themselves).

The novelty of this work can be summarized as follows:

- A unified Agent design and processing framework basing on a reactive activity-transition agent behaviour and programming model. Agent interaction is provided by tuple spaces and signal propagation between agents.
- Stack based Virtual Machines (SVM) are used to execute optimized program code embedding the agent behaviour, data and control state in code frames
- The SVM is operating system independent and can be implemented directly in hardware and software including JavaScript
- The JavaScript implementation of the SVM enables the integration of sensor networks and agent-based sensor and information processing in the Internet and Intranet domains.
- The SVM can be embedded in *HTML* content and turns a browser in an agent processing platform.
- A object-capability-based Remote Procedure Call (RPC) communication interface and a distributed graph-linked broker service enables the deployment of client-side applications like browser as agent processing platforms.

II. The State-based Reactive Agent Behaviour Model and AAPL Programming Language

The agent model summarized in this section (for details see **[1][3][4]**) bases on the mobile processes model introduced by Milner **[16]** several decades ago. An agent can be considered as a computational unit situated in an environment and world, which performs computation, basically hidden for the environment, and interacts with the environment to exchange basically data. A common computer is specialised to the task of calculation, and interaction with other machines is encapsulated by calculation and performed traditionally by using messages. An agent behaviour can be reactive or proactive, and it has a social ability to communicate, cooperate, and negotiate with other agents. Proactiveness is closely related to goal-directed behaviour including estimation and intentional capabilities.

II-A. Activity-Transition Graphs

The behaviour of an activity-based agent is characterized by an agent state, which is changed by activities. Activities perform perception, plan actions, and execute actions modifying the control and data state of the agent. Activities and transitions between activities are represented by an activity-transition graph (ATG). The transitions start activities commonly depending on the evaluation of agent data (body variables), representing the data state of the agent. The ATG behaviour model is fundamental for Activity-based Agent Programming Language (*AAPL*).

An activity-transition graph, related to the agent classes, discussed later, consists of a set of activities $\boldsymbol{A}=\{A_1,A_2,..\}$, and a set of transitions $\boldsymbol{T}=\{T_1\ (C_1),T_2\ (C_2),..\}$, which represent the edges of the directed graph. The execution of an activity, composed itself of a sequence of actions and computations, is related with achieving a sub-goal or a satisfying a prerequisite to achieve a particular goal, e.g., sensor data processing and distributions.

Usually agents are used to decompose complex tasks in simpler ones. Agents can change their behaviour based on learning and environmental changes, or by executing a particular sub-task with only a sub-set of the original agent behaviour.

An ATG describes the complete agent behaviour. Any sub-graph and part of the ATG can be assigned to a subclass behaviour of an agent. Therefore modifying the set of activities $\boldsymbol{A}$ and transitions $\boldsymbol{T}$ of the original ATG introduces several sub-behaviour for implementing algorithms to satisfy a diversity of different goals. The reconfiguration of activities $\underline{\boldsymbol{A}} = \{ \boldsymbol{A}_1 \subseteq \boldsymbol{A}, \boldsymbol{A}_2 \subseteq \boldsymbol{A}, ..\}$ from the original set $\boldsymbol{A}$ and the modification or reconfiguration of transitions $\underline{\boldsymbol{T}} = \{ \boldsymbol{T}_1, \boldsymbol{T}_2, ..\}$ create dynamic supersets of ATGs and enable agent sub-classing at run-time.

II-B. The Activity-based Agent Programming Language (AAPL)

The AAPL programming model should optimally match the requirements of MAS deployed in unreliable sensor and wide-area distributed networks, keeping low-resource nodes with low computational power in mind. On one hand, *AAPL* should reflect the core concepts of agents, on the other hand *AAPL* should provide core concepts of traditional programming language to ease the programming of widely used algorithms.

The agent behaviour, perception, reasoning, and the action on the environment are encapsulated in agent classes, with activities representing the control state of the agent reasoning engine, and conditional transitions connecting and enabling activities. Activities provide a procedural agent processing by a sequential execution of imperative data processing and control statements. Agents can be instantiated by other agents from a specific class at run-time. A multi-agent system composed of different agent classes enables the factorization of an overall global task in sub-tasks, with the objective of decomposing the resolution of a large problem into agents in which they communicate and cooperate with one other.

AAPL supports the following statements and constructors:

- Agent Class Definition consisting of body variables, activities, transitions, handlers, and common functions;
- Computational and control flow statements: assignment, branches, loops, exception handling;
- Cooperation and Communication with tuple spaces and signal messages (carrying simple data);
- Agent instantiation from agent classes, forking, destroying;
- Agent mobility by migration;
- Agent behaviour modification (e.g., ATG reconfiguration).

II-C. Multi-Agent Interaction

In parallel and distributed systems the communication, synchronization, and data exchange of a collection of data processing units (processes or agents) gains significant importance. A common approach for parallel systems is a shared memory based communication paradigm, but which generates a high computational dependency of the processing units among themselves and regarding the platform. Loosely coupled distributed systems like MAS require a different communication strategy.

Tuple-Spaces. One well known and common distributed interaction model is the tuple-space. Agents can communicate with each other by accessing a tuple space database service available on each network node and that is provided by the agent processing platform (a node in the Bigraph model, see bottom of Fig **1**), used for synchronized data exchange among a collection of individual agents, which was proposed in **[18]** and **[19]** as a suitable MAS interaction and coordination paradigm.. A tuple space is a logically shared memory and is used for synchronized data exchange between producer and consumer, a common approach for solving communication problems of loosely coupled autonomous or semi-autonomous processing units. Tuple spaces are generative, which means a tuple can survive the creator beyond its lifetime. The scope and visibility of a tuple space database can be unlimited and visible and distributed in the whole network, or limited to a local scope, e.g., network node level. A tuple space provides abstraction from the underlying platform architecture, and offers a high degree of platform independency, vital in a heterogeneous network environment.

For the sake of simplicity the scope of a tuple space can be limited to the node boundary, such that there are multiple tuple spaces distributed in the network. Information can be carried by mobile agents between nodes. A tuple space communication model has the advantage of shielding the underlying node and agent processing platform. Access of tuple spaces require only a small set of simple operations {out, in, rd, in?, rd?, rm, eval}, which transfer tuples between a producer or consumer and the database. Since tuples consist of type-tagged values and patterns the tuple space communication is type-safe and strong computational bindings can be avoided.

AAPL Agents. In the Bigraph model *AAPL* agents have different ports. One static port is the platform link, required to execute an agent process. Another port is used for the linking of an agent with a tuple-space (#=1). An *AAPL* agent can have only one tuple-space access and link at any time maximal. The propagation of signals introduce further ports and dynamic links to other agents (#=0..n), see Fig. **1**. The communication links introduce virtual domains, in Fig. **1** these are the agent groups $\{A_2, A_3, A_4\}$ and $\{A_5, A_6\}$. These virtual domains are dynamic, regarding the spatial location and extension, and the agents which are part of the virtual domain. Often agent parent-child trees spawn the virtual domains using signal interaction, but agents of initially different virtual domains can interact by using the tuple-spaces, extending and merging different virtual domains. The spatial extension of virtual MAS domains is constrained by the connectivity graph of the processing nodes.

Signal propagation from a source to a destination agent requires the connectivity of nodes if the agents are executed on spatially different nodes. Tuples stored in tuple-spaces are persistent. That means a tuple *t*, which was produced by an agent Ag_1 and stored in a tuple-space TS_1, and agent Ag_1 is finally migrating to another node location, can be consumed by a different agent Ag_2, now having a historical relation and link to the other agent Ag_1.

Signals. In contrast, *signals*, which can carry additional scalar data values, can be used for local (in terms of the node scope) and global (in terms of the network scope) domain agent interaction. In contrast to the anonymous tuple-space interaction, signals are directly addressed to a specific agent or a group of agents. The delivery of signals is not reliable in the case the agents raising and receiving the signal are not processed on the same node. An agent being ready to receive signals has to provide a signal handler for this signal, a function that is executed asynchronously to the agent ATG execution.

III. The Agent Code Processing Platform

In this work, the agents are implemented with Agent Forth program code that is executed on virtual stack machines, which can be implemented alternatively on hardware (System-on-Chip), simulation, and software level, which can be embedded in microcontroller, desktop applications, web applications, or server programs. The agent program code (see **[1]**) is a self-containing and self-initializing unit embedding the (private) agent data and the current control state of the agent, which simplifies migration significantly. This machine program is encapsulated in code frames with a specific layout. The program is able to modify itself by using code morphing, leading to a low computational dependency from the current execution environment, which is vital to strong heterogeneous environments. There is only a small set of knowledge about the program which is required by the VM to execute the agent program, and vice versa. Migration of agents requires only the transfer of the code frame from one platform to another. The data and control state of an agent program is stored in the code frame, too. There are two different *Agent FORTH* levels, one supporting high-level constructs like loops and branches (*AFL*), and one low-level machine subset (*AML*) that can be directly executed by the *AFVM* platform. *AFL* has similar operational semantics than *AAPL*. Thus the *AAPL* agent class behaviour definition can be directly compiled to the *AFL* level, finally compiled to *AML* with a specific code frame layout.

In **[2]** and **[3]** there is an example for the *AAPL* behaviour model of a simple explorer agent that is sent out from an agent on a specific network node. The explorer agent has the goal to find another node having a specific feature that is stored in the (local) tuple space database. If the explorer agent found the feature (activity check), it will return the original root node and stores the feature in the tuple space with the relative delta position of the node where the feature

tuple was found (activity deliver). The explorer agent moves through the network in a random direction until a maximal number of hop counts is reached (parameter radius, activity migrate). The respective *AFL* program (see **[1]**) reflects roughly the operational semantics and structure of the *AAPL* program source. The compiled *AML* machine program that can be executed by the *AVM* consists of a boot section at the beginning of the code frame, followed by a data section storing the private agent variables and parameters. Finally all activities and the transition table conclude. The entire machine program requires less than 400 words (800 bytes for a 16 Bit machine), which can be efficiently transferred between different processing hosts.

III-A. AFVM Platform Architecture

The virtual machine (*AFVM*, discussed in depth in **[1]**) executing tasks bases on a traditional *FORTH* stack processor architecture and an extended zero-operand word instruction set (αFORTH). Most instructions operate directly on the data stack *DS* and the control return stack *RS*. A code segment *CS* stores the program code with embedded data. The program is mainly organized by a composition of words (functions). A word is executed by transferring the program control to the entry point in the *CS*; arguments and computation results are passed only by the stack(s). There are multiple virtual machines, each attached to (private) stack and code segments. There is one global code segment *CCS* storing global available functions and code templates which can be accessed by all programs. A dictionary is used to resolve *CCS* code addresses of global functions and templates.

The program code frame of an agent is a standalone and auto-initializing unit that encapsulates basically four parts: 1. A look-up table and embedded agent body variable definitions, 2. Word definitions defining agent activities and signal handlers (procedures without arguments and return values) and generic functions, 3. Bootstrap instructions for the setup of agents in a new environment (i.e., after migration or on first run), and 4. The transition table calling activity words and branching to succeeding activity transition rows depending on the evaluation of conditional computations with private data (variables). The transition table section can be modified by the agent by using special instructions. Code morphing can be applied to the currently executed code frame or to any other code frame of the VM.

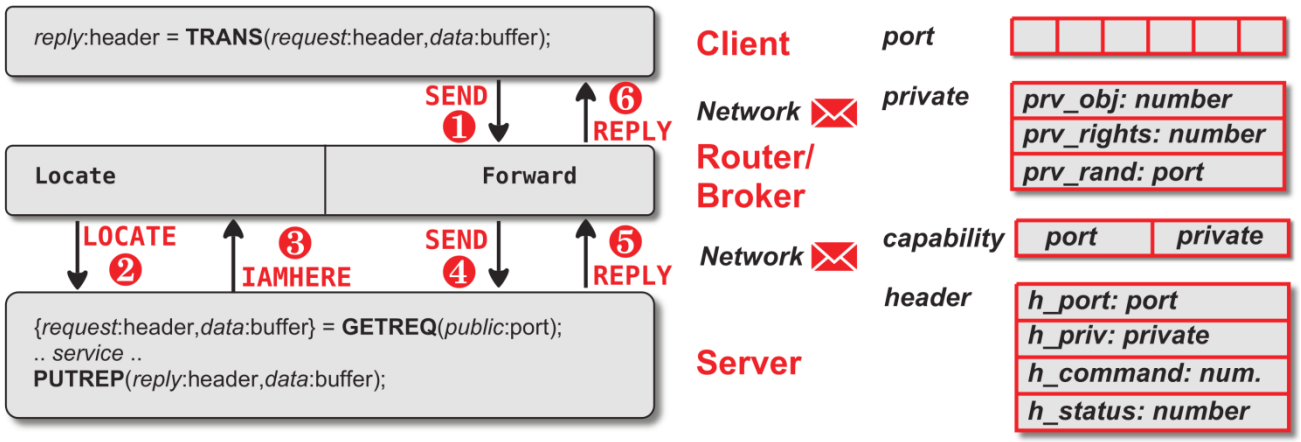
Each VM processor is connected with an agent process manager (AM). The VM and the agent manager share the same VM code segment and the process table (PT). The process table contains only basic information about processes required for the process execution.

Commonly the number of agent tasks N_A executed on a node is much larger than the number of available virtual machines N_V. Thus, efficient and well-balanced multi-task scheduling is required to get proper response times of individual agents. To provide fine grained granularity of task scheduling, a token based pipelined task processing architecture was chosen. A task of an agent program is assigned to a token holding the task identifier of the agent program to be executed. The token is stored in a queue and consumed by the virtual machine from the queue. After a (top-level) word was executed, leaving an empty data and return stack, the token is either passed back to the processing queue or to another queue (e.g., of the agent manager). Therefore, the return from an agent activity word execution (leaving empty stacks) is an appropriate task scheduling point for a different task waiting in the VM processing token queue. This task scheduling policy allows fair and low-latency multi-agent processing with fine grained scheduling. Furthermore, this kind of task scheduling enables the *JavaScript* implementation, discussed in Sec. IV-E.

IV. The JavaScript Web Platform JAVM

The mobility of agents is handled basically by the agents themselves, and there is no advanced routing provides by the platform. They make decisions about the migration direction and the selection of neighbour nodes, usually basing on some geometrical structures given by the network topology. For example, a material-integrated sensor network embedded in a wind energy wing used for Load Monitoring has a mesh-like network topology consisting of nodes that are connected with their nearest neighbours. Delivering of sensor data to dedicated computing nodes can be performed simply by travelling to the outside of the network and by searching. In the Internet context this geometrical structure and the neighbourhood connectivity do not exist, or at least they are not visible, increasing the decision and reducing the knowledge space of agents significantly. First of all, the migration decision of agents must base on different features and knowledge. Furthermore, the Internet consists of two different kinds of network nodes: Nodes capable of providing a public visible service, called servers, and nodes that cannot publish server ports. But in distributed systems each node must be capable of offering services. Two computers can only connect if at least one computer has public server ports, otherwise an external brokerage service is required. Web browsers are usually processed on client computer nodes and are not visible in the network. Therefore, agents can't select a client-interface-only node or process for migration directly and autonomously due to the missing visibility in the communication network, as this is the case in traditional sensor or embedded networks.

Two main issues arising in Internet applications using mobile agents must be addressed: 1. The definition and the knowledge representation of virtual/artificial neighbourhood connectivity in loosely coupled and hierarchical graph-based networks based on semantic rather on physical connectivity. 2. The visibility and deployment of pure client-side applications like Web browsers and computers hidden in private or restricted networks as agent processing platforms capable of receiving, processing, and sending of agents.



Def. 2 RPC-based client-server communication types, operations, and protocol schema (phases of a transaction)

To enable the distributed agent processing in browser and applications running on generic computers connected by the Internet, the previously introduced *Agent Forth Virtual Machine* (*AFVM*) platform was implemented in JavaScript that can be executed either by a *node.js* interpreter or by any browser capable to execute *JavaScript* code. The *AFVM* was integrated in a distributed operating system layer, also implemented entirely in *JavaScript*, discussed in the following subsections, composing the *JAVM* platform. The transition from peer-to-peer networks to routed and hierarchical networks like the Internet requires some methodological and architectural changes, introducing the aforementioned broker service, discussed below.

IV-A. Inter-Node Communication and RPC

Nodes offering agent processing capabilities connected in the Internet domain usually not communicating peer-to-peer like in sensor networks with mesh topologies. Instead routing is used to establish communication between different application processes executed on nodes probably located far away. One well known inter-process communication approach is the Remote Procedure Call (RPC), e.g., extensively used in the distributed operating system Amoeba **[21]**, or on the top of existing operating system, e.g., offered by the distributed Common Object Request Broker Architecture (*CORBA*) framework. The capability-based RPC communication from the Amoeba OS was already successfully implemented in VM environments executed on top of existing operating systems (*VAMNET*, **[5]**).

The RPC communication interface is used in this work for the inter-platform communication, e.g., for transferring agent program code to another platform or to access distributed file and naming services. The RPC ontology consists of servers and clients communicating by using a set of operations. A server performs a GETREQ operation to publish a listening on a public server port, and a client performs a transaction TRANS operation to access a server identified by the public server port. Each server handles a set of objects, identified by capabilities that are tuples $\langle port, obj, rights, rand \rangle$, consisting of the server port, an object number, a rights field, and a private protection field authorizing the rights field. A transaction operation transfers object capabilities to the server that handles the request and finally replies by using the PUTREP operation. Therefore, a client transaction is synchronous and blocks the client process until the reply arrives or an error occurred (time-out). The localization of the server and the routing of the messages is hidden by the RPC layer, or more precisely by the underlying protocol layer, shown in Def. **1**. The localization is basically performed by broad- or multicasting LOCATE messages to nodes in the current domain and finally to a limited number of boundary domains. Each node monitors the locally registered servers, and replies with a IAMHERE message. Nodes are identified with ports, too.

The RPC communication is encapsulated in *HTTP* messages with *XML* content and transferred using the generic *HTTP* protocol, discussed in section IV-C. The RPC header and data is stored inside *XML* tags with compacted hexadecimal coded text, on one hand complaining with the *XML* standard, on the other hand reducing and optimizing the payload. The binary byte data is coded with two hexadecimal digits for each data byte. Each RPC server (process) can act as a client, too, and vice versa.

IV-B. Domains as Organization Structures and the Directory Name Service

Domains are groups of agent processing nodes that are coupled in a network. Agents can migrate between nodes of a group. A node can be assigned to more than one domain, enabling the migration of agents between domains. Node domain composition bases on

1. Geometrical localization and proximity, basically expressing and simulating neighbourhood connectivity
2. Information and data context
3. Tasks to be performed, cooperative goals to be satisfied
4. Logical network domains

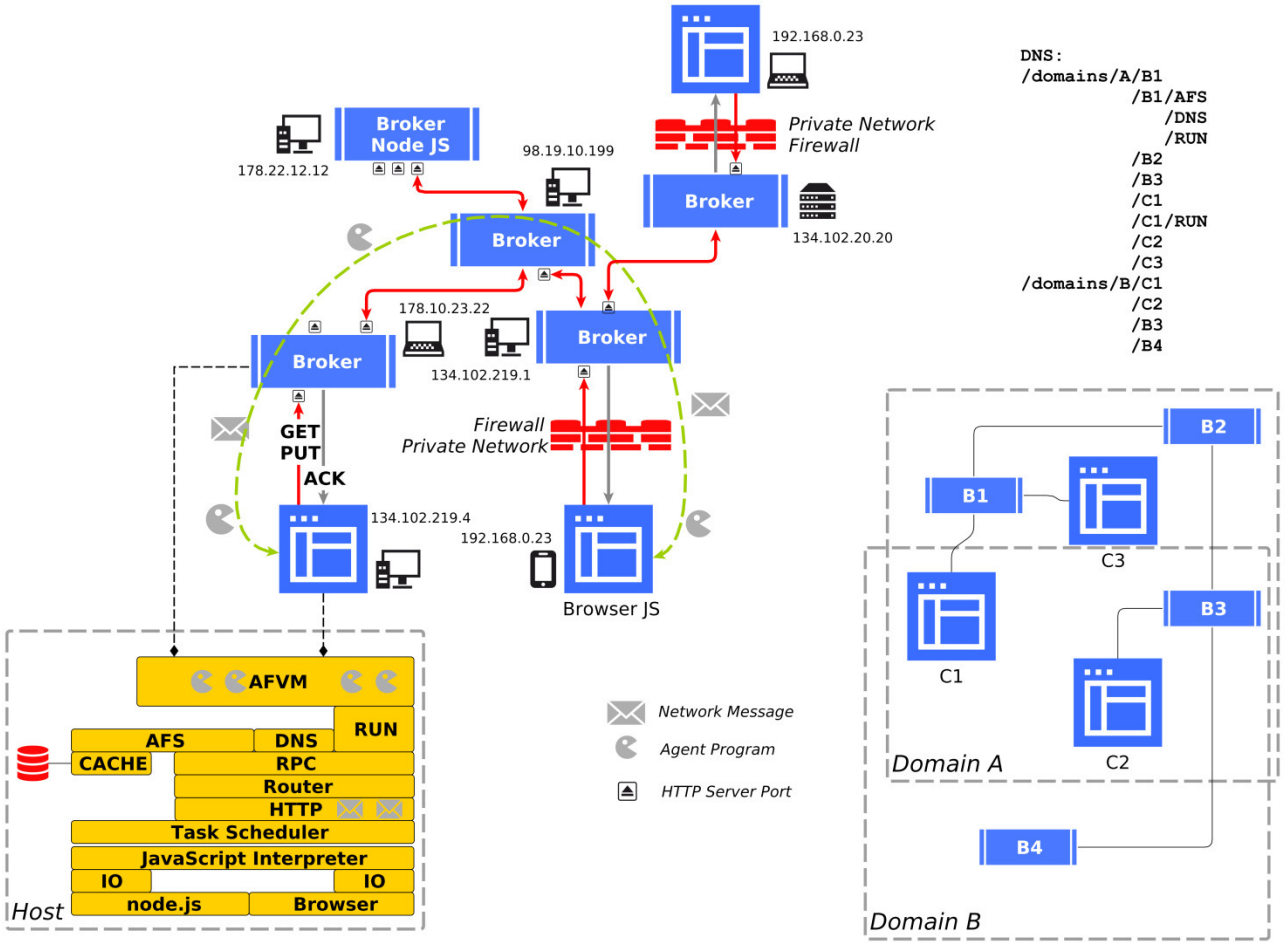


Fig. 2 (Left, Centre) Broker Network with HTTP server ports and client applications (browser, node.js client-side) connecting to the public visible broker server ports. Client-to-Client communications takes place over the broker servers. (Left, Bottom) The JavaScript agent platform JAVM and the modules and services available on each host (Right) Different nodes can be bound to (overlapping) domains published in the DNS.

Domains can be expressed by paths similar to directory trees that are handled usually by a file system. In this work a distributed and unified Directory Name Service (*DNS*) us used that provides a database to publish (capability-name) pairs organized in trees. Each object in the distributed system is related to a capability, which is serviced by a specific server. For example, a file containing the agent program code is serviced by a file server. A directory containing domains is an object, too, handled by the *DNS* server. An agent platform that processes agents programs is another kind of object, handled by a run server that exists on each node. Agents are objects in this sense, but they don't belong to a specific server, therefore they are handled as mobile and autonomous severs. In Fig. **2**, an example for a composition of domains consisting of network nodes that are not directly connected is shown.

IV-C. Broker Service

The integration and network connectivity of client-side application programs like Web browsers as an active agent processing platform requires client-to-client communication capabilities, which is offered in this work by a broker server that is visible in the Internet or Intranet domains. Though there are already some approaches for interconnecting browser applications directly (client-to-client communication using *WebSockets* or *WebRTC* **[20]** and *HTML5* standards), they are not supported by all browsers and require some external server for the connection brokerage, too. Furthermore, *WebSockets* are still under development and there are many browser incompatibilities. To provide compatibility with and among all existing browser applications none of these technologies were used. Instead, an object-capability-based RPC inter-process communication with a broker server operating as a router was invented. Client applications communicate with the broker by using the generic *HTTP* client protocol and the GET and PUT operations. RPC messages are encapsulated in *HTTP* requests. If there is a RPC server request passed to the broker, the broker will cache the request until another client-side host performs a matching transaction to this server port. The transaction is passed to the original RPC server host in the reply of a *HTTP* GET operation.

But the deployment of one central broker server introduces a single-point-of-failure and is limiting the communication bandwidth and the scaling capability significantly. To overcome these limitations, a hierarchical broker server network is used. Each broker in this broker graph can be the root of a sub-graph and can be a service end-point

(i.e., providing directory and name services), a router between clients and other broker servers, and an interface bridge to a non-IP based network, e.g., a sensor network. A broker is just an application program capable of running on any computer visible globally in the Internet or more locally in some Intranet domains.

An agent processing node (e.g., a host application) that cannot publish IP server ports must connect to one of the broker servers visible in the network. Usually this should be a server located nearby. Each node is associated with a host port that is communicated to the broker server now handling and forwarding service requests for this specific host, shown on the lower left side of Fig. **2**. Each client-side host collects periodically pending and queued service request messages (or replies of services requests) from the broker server and passes services replies back to the broker server that forwards the reply to the appropriate host performing originally a transaction. If the two hosts involved in a RPC transaction are not handled by the same broker server, the source broker server must forward request and reply messages to the appropriate destination broker server, shown in Fig. **2** by the green dotted path line. Furthermore, a broker server must handle local RPC transactions and local RPC servers and, too.

IV-D. The Node Service Platform

In addition to the services provides by the agent processing platform (i.e., the agent manager and the tuple-space database), each network broker node and optionally each browser or client-side application provide a file system service (Atomic File System Service *AFS*), the aforementioned Directory and Naming Service (*DNS*), and a run server connected to the agent processing platform (required on each host). The run server provides the public port for agent execution, migration, and signal message propagation between agents.

IV-E. The JavaScript Implementation

There are basically two different execution environments for the execution of *JavaScript* (*JS*) programs: The server-side standalone *node.js* interpreter and the client-side *JS* interpreter embedded in browser applications. The *node.js* interpreter can execute a *JS* program directly (with source-to-machine code compilation on demand), whereas the browser executes *JS* embedded in *HTML* content only. There are *node.js* modules enabling the setup of *HTTP* servers, modules for accessing files on the local file system, and many more OS related programming interfaces not available in the client-side browser *JS*.

The implementation of the entire network node services, the RPC communication, and the agent processing platform with JavaScript is a challenge, but offers significant advantages with respect to portability, compatibility, and the design unification for server-side and client-side-only platforms (e.g., browsers). The basic modules implemented on each host (and browser application) are shown on the left bottom side of Fig. **2**, consisting at least of the RPC module, the *HTML* wrapper, and the agent processing platform *AFVM*.

JavaScript is executed strictly single threaded, though functions can be executed in parallel and concurrently, there is no concept of process blocking or any other synchronization. In JavaScript programs input-output operations are mainly performed with asynchronous callback functions. But all RPC services, the agent processing platform, and servers operate inherently multi-threaded and synchronously.

To overcome this execution limitation, a *Task Scheduler* (TSCH) was invented that simulates parallel multiprocess execution and enables virtual process blocking for the synchronization of processes. Each process consists of a set of activities (functions) that are enabled by a conditional transition expression (that can be a constant true value). The scheduler executes all activity functions sequentially that have a satisfied transition condition. Blocking of a process sets a process specific blocking variable (the guard GD) that is part of the transition condition from the blocking activity to the next one to be executed after the process was woken up again. Furthermore, there are block, conditional, and loop scheduling constructors easing the programming of processes. All RPC operations are prepared for the scheduler management. Though callback functions are still used, a single program flow of processes can be constructed on programming level.

The client-side Browser *JS* implementation is created by compacting and relocating server-side dependencies (using *browserify*, *envify*, and *uglifyjs* for minimizing), and requires typically about 500kB text size.

V. Use-Case: Cloud Based Adaptive Manufacturing and Robots as Products

This section outlines a big application use-case for the introduced agent processing platform with an architecture for additive and adaptive manufacturing based on a closed-loop sensor processing approach, extended with data mining concepts combined with Internet-of-thing architectures. Additive and adaptive cloud-based design and manufacturing are attractive in the field of robotics, not only limited to industrial production robotics, mainly targeting service robots and semi-autonomous carrier robots. In cloud-based manufacturing, the consumer of the products is integrated in the cloud-based manufacturing process **[6]**, directly involved in the manufacturing process using distributed cloud computing and distributed storage solutions.

Robots can be considered as active, mobile, and autonomous data processing units that are commonly already connected to computer networks and infrastructures. Robots use inherent sensing capabilities for their control and task satisfaction, commonly using integrated sensing networks with sensor preprocessing, deriving some inner state of the robot, e.g., mechanical loads applied to structures of the robot or operational parameters like motor power and temperature. The availability of the inner perception information of robots enable the estimation of working and health conditions initially not fully considered at design time. The next layer in cloud-based adaptive manufacturing process can be the inclusion of the products themselves

delivering operational feedback to the current design and manufacturing process, leading to a closed-loop evolving design and manufacturing process with an evolutionary touch, shown in Fig. **3**. This evolutionary process adapts the product design, e.g. the mechanical construction, for future product manufacturing processes based on a back propagation of the perception information (i.e., recorded load histories, working and health conditions of the product) collected by living systems at run-time. The currently deployed and running series of the product enhances future series, but not in the traditional coarse-grained discrete series iteration. This process can be considered as a continuously evolving improvement of the robot by refining and adapting design parameters and constraints that are immediately migrated to the manufacturing process. A robot consists of a broad range of parts, most of them are critical for system failures. The most prominent failures are related to mechanical and electro-mechanical components, which are caused by overload conditions at run-time under real conditions not to be considered or unknown at initial design time.

The integration of robots as product and their condition monitoring in a closed-loop design and manufacturing process is a challenge and introduces distributed computing and data distribution in strong heterogeneous processing and network environments. One major question to be answered is the sensing of meaningful condensed product condition information and the delivery to the designer and factory. The proposed mobile agent model offers a self-contained and autonomous virtual processing unit that is well suited for such large-scale applications. The mobile agents represent mobile computational processes that can migrate in the Internet domain and as well in sensor networks.

Agents are already deployed successfully for scheduling tasks in production and manufacturing processes **[7]**, and newer trends poses the suitability of distributed agent-based systems for the control of manufacturing processes **[8]**, facing not only manufacturing, but maintenance, evolvable assembly systems, quality control, and energy management aspects, finally introducing the paradigm of industrial agents meeting the requirements of modern industrial applications. The MAS paradigm offers a unified data processing and communication model suitable to be employed in the design, the manufacturing, logistics, and the products themselves.

The scalability of complex industrial applications using such large-scale cloud-based and wide area distributed networks deals with systems deploying thousands up to million agents. But the majority of current laboratory prototypes of MAS deal with less than 1000 agents **[8]**. Currently, many traditional processing platforms cannot yet handle big numbers with the robustness and efficiency required by industry **[9][10]**. In the past decade the capabilities and the scalability of agent-based systems have increased substantially, especially addressing efficient processing of mobile agents.

There programmable agent processing platform introduced in this work can be deployed in strong heterogeneous network environments, ranging form single microchip up to WEB *JavaScript* implementations, all being fully compatible on operational and interface level, and hence agents can migrate between these different platforms. Multi-agent systems can be successfully deployed in sensing applications, e.g., structural load and health monitoring, with a partition in off- and online computations **[2]**. Distributed data mining and Map-Reduce algorithms are well suited for self-organizing MAS. Cloud-based computing, as a base for cloud-based manufacturing, means the virtualization of resources, i.e., storage, processing platforms, sensing data or generic information.

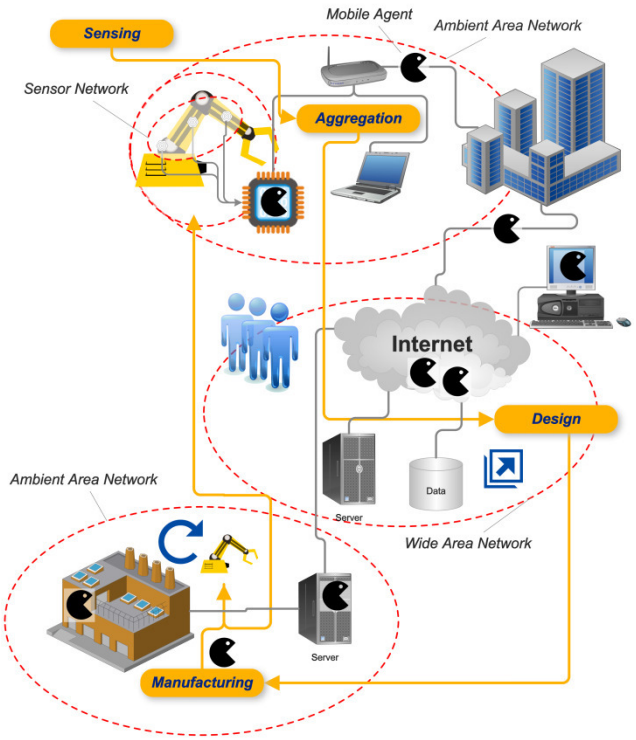


Fig. 3 Additive and adaptive Manufacturing with back propagation of sensing data using mobile *AAPL* agents and the JAVM/PAVM platform.

Traditional closed-loop processes request data from sources (products, robots) by using continuos request-reply message streams. This approach leads to a significant large amount of data and communication activity in large-scale networks. Event-based sensor data and information distribution from the sources of sensing events, triggered by the data sources (the robots) themselves, can improve and reduce the allocation of computational, storage, and communication resources significantly.

A cloud in terms of data processing and computation is characterized by and composed of: 1. A parallel and distributed system architecture; 2. A collection of interconnected virtualized computing entities that are dynamically provisioned; 3. A unified computing environment and unified computing resources based on a service-level architecture; 4. A dynamic reconfiguration capability of the virtualized resources (computing, storage, connectivity and networks).

Cloud-based design and manufacturing is composed of knowledge management, collaborative design, and distributed manufacturing. Adaptive design and manufacturing enhanced with perception delivered by the products incorpo-

rates finally the products in the cloud-based design and manufacturing process.

Agent Classes. Different agent classes are defined that satisfy different sub-goals: event-based sensor acquisition including sensor fusion (Sensing), aggregation and distribution of data, preprocessing of data and information mapping, search of information sources and sinks, information delivery to databases, delivery of sensing, design, and manufacturing information, propagation of new design data to and notification of manufacturing processes, notification of designer, end users, update of models and design parameters. Most of the agents can be transferred in messages with a size lower than 4kB.

VI. Conclusion and Outlook

In this work, a novel **Agent Processing Platform** architecture for code-based mobile agents in large-scale and wide-area heterogeneous networks including low-resource microchip nodes embedded in sensor networks was introduced. The standalone agent processing platform, a multi-core stack processor, can be implemented entirely on microchip level, and requires no operating system and no boot code. Alternatively, the processing platform can be implemented efficiently in software with code and operational compatibility, enabling the deployment in heterogeneous network environments, inter-connecting hardware and software platforms executed on generic microprocessors. The *JavaScript* implementation of the processing platform together with a minimal distributed operating layer consisting of a broker, RPC, run, file, and naming services enables the integration of body area, ambient, and sensor networks in the Internet domain, a prerequisite for the future of Internet-of-Things and Sensor Clouds in daily use computing environments. Agents can migrate between different hardware and software platforms (they are compatible on the execution level) by migrating the program code of the agent, embedding the state and the data of an agent, too. A broker service enables the integration of hosts (generic computers, mobile devices, ...) that are not visible in the Inter- or Intranet domains and that cannot publish server communication ports.

Using this broker service, which is composed of a graph-based network of single broker server applications, each computing device capable of executing *JavaScript* code can act as an agent processing platform. This agent processing platform is capable of receiving mobile agents form other platforms and hosts. The broker service creates virtual connectivity based on domains.

References

[1] S. Bosse, *Design and Simulation of Material-integrated Distributed Sensor Processing with a Code-based Agent Platform and mobile Multi-Agent Systems*, MDPI Sensors, 2015 (2), pp. 4513-4549, 2015, http://dx.doi.org/10.3390/s150204513

[2] S. Bosse and A. Lechleiter, *Structural Health and Load Monitoring with Material-embedded Sensor Networks and Self-organizing Multi-agent Systems*, Procedia Technology, Proceeding of the 2nd SysInt Conference, Bremen, Germany, 2014, http://dx.doi.org/10.1016/j.protcy.2014.09.039

[3] S. Bosse, *Distributed Agent-based Computing in Material-Embedded Sensor Network Systems with the Agent-on-Chip Architecture*, IEEE Sensors Journal, http://dx.doi.org/10.1109/JSEN.2014.2301938

[4] S. Bosse, *Design of Material-integrated Distributed Data Processing Platforms with Mobile Multi-Agent Systems in Heterogeneous Networks*, ICAART 2014, http://dx.doi.org/10.5220/0004817500690080

[5] S. Bosse, *VAMNET: the Functional Approach to Distributed Programming*, SIGOPS Oper. Syst. Rev., 40, pp. 108-114, 2006, http://dx.doi.org/10.1145/1151374.1151378.

[6] D. Wu, J. L. Thames, D. W. Rosen, and Dirk Schaefer, *Towards A Cloud-based Design and Manufacturing Paradigm: Looking Backward, Looking Forward*, in Proceedings of the ASME 2012 International Design Engineering Technical Conference & Computers and Information in Engineering Conference, IDETC/CIE 2012 August 12-15, 2012, Chicago, Illinois, USA, 2012

[7] M. Caridi and A. Sianesi, *Multi-agent systems in production planning and control: An application to the scheduling of mixed-model assembly lines*, Int. J. Production Economics, vol. 68, pp. 29–42, 2000.

[8] P. Leitão and S. Karnouskos (ed.), in *Industrial Agents Emerging Applications of Software Agents in Industry*. Elsevier, 2015.

[9] V. Maŕık, and D.C. McFarlane, 2005. *Industrial adoption of agent-based technologies*. IEEE Intell. Syst. 20 (1), 27–35.

[10] M. Pechoucek, and V., Maŕık, 2008. *Industrial deployment of multi-agent technologies: review and selected case studies*. Auton. Agent. Multi-Agent Syst. 17 (3), 397–431.

[11] M. Guijarro, R. Fuentes-fernández, and G. Pajares, *A Multi-Agent System Architecture for Sensor Networks*, Multi-Agent Systems - Modeling, Control, Prog., Simulations and Applications, 2008.

[12] A. Rogers, D. D. Corkill, and N. R. Jennings, *Agent Technologies for Sensor Networks*, IEEE Intelligent Systems, vol. 24, no. 2, 2009.

[13] X. Zhao, S. Yuan, Z. Yu, W. Ye, and J. Cao, *Designing strategy for multi-agent system based large structural health monitoring*, Expert Systems with Applications, 2008, 34(2), 1154–1168. doi:10.1016/j.eswa.2006.12.022

[14] J. Liu, *Autonomous Agents and Multi-Agent Systems,* World Scientific Publishing, 2001 (ISBN 981-02-4282-4)

[15] R. Milner, *The space and motion of communicating agents*. Cambridge University Press, 2009.

[16] R. Milner, *Communicating and mobile systems: the π-calculus*, Cambridge University Press, Cambridge (1999)

[17] L. Cardelli and A: Gordon, *Mobile Ambients*. Theoretical Computer Science, Special Issue on Coordination 240(1), 177–213 (2000)

[18] L. Chunlina, L. Zhengdinga, L. Layuanb, and Z. Shuzhia, *A mobile agent platform based on tuple space coordination, Advances in Engineering Software*, vol. 33, no. 4, pp. 215–225, 2002

[19] Z. Qin, J. Xing, and J. Zhang, *A Replication-Based Distribution Approach for Tuple Space-Based Collaboration of Heterogeneous Agents*, Research Journal of Information Technology, vol. 2, no. 4. pp. 201–214, 2010

[20] S. Loreto and S. Pietro Romano, *Real-time communications in the web: Issues, achievements, and ongoing standardization efforts*, IEEE Internet Computing, vol. 16, no. 5, pp. 68–73, 2012.

[21] S. J. Mullender and G. van Rossum, *Amoeba: A Distributed Operating System for the 1990s*, IEEE Computer, vol. 23, no. 5, pp. 44–53, 1990.

9th International Workshop on Multi-Agent Systems and Simulation

MULTI-AGENT systems (MASs) provide powerful models for representing both real-world systems and applications with an appropriate degree of complexity and dynamics. Several research and industrial experiences have already shown that the use of MASs offers advantages in a wide range of application domains (e.g. financial, economic, social, logistic, chemical, engineering). When MASs represent software applications to be effectively delivered, they need to be validated and evaluated before their deployment and execution, thus methodologies that support validation and evaluation through simulation of the MAS under development are highly required. In other emerging areas (e.g. ACE, ACF), MASs are designed for representing systems at different levels of complexity through the use of autonomous, goal-driven and interacting entities organized into societies which exhibit emergent properties The agent-based model of a system can then be executed to simulate the behavior of the complete system so that knowledge of the behaviors of the entities (micro-level) produce an understanding of the overall outcome at the system-level (macro-level). In both cases (MASs as software applications and MASs as models for the analysis of complex systems), simulation plays a crucial role that needs to be further investigated.

Topics

MAS&S'15 aims at providing a forum for discussing recent advances in Engineering Complex Systems by exploiting Agent-Based Modeling and Simulation. In particular, the areas of interest are the following (although this list should not be considered as exclusive):

- Agent-based simulation techniques and methodologies
- Discrete-event simulation of Multi-Agent Systems
- Simulation as validation tool for the development process of MAS
- Agent-oriented methodologies incorporating simulation tools
- MAS simulation driven by formal models
- MAS simulation toolkits and frameworks
- Testing vs. simulation of MAS
- Industrial case studies based on MAS and simulation/testing
- Agent-based Modeling and Simulation (ABMS)
- Agent Computational Economics (ACE)
- Agent Computational Finance (ACF)
- Agent-based simulation of networked systems
- Scalability in agent-based simulation

Position Papers of the Federated Conference on
Computer Science and Information Systems pp. 249–256

DOI: 10.15439/2015F417
ACSIS, Vol. 6

Multi-agent simulation of the world found in the G. R. R. Martin's novel "Sandkings"

Jakub Ciecierski, Viet Ba Mai, Michał Słupczyński and Wojciech Zyskowski
Faculty of Mathematics and Information Science, Warsaw University of Technology
Plac Politechniki 1, 00-660 Warsaw, Poland

***Abstract*—George R. R. Martin's novel entitled *Sandkings* introduces tribes of ant-like creatures which, when locked in a terrarium deprived of food, fight for survival. The tribes differ in the color of their armor. However, in the novel, despite the extreme conditions, they've never attempted to kill the queens of different fractions for *cannibalistic* purposes. Approaching the situation from the perspective of profit-driven logic, lack of such behaviour is highly questionable. In multiple cases, attacking another Maw to eat her could improve the odds of survival in a hostile environment.**

To provide an initial attempt to answer the question of whether or not the queens should order an attack and try to "eat each other", we have developed a multi-agent simulation based on the novel. Through analysis of its results we hoped to prove hypothesis that with the increased hostility of environment the *Sandkings* would develop cannibalistic behaviours and fight until eventually only one tribe is left.

I. Introduction

SAND KINGS is a science-fiction novel written by George R.R. Martin in 1979 [1]. There, Simon Kress, the main protagonist of the story, is as a collector of lethally dangerous, exotic, and mostly extraterrestrial, animals. Due to his prolonged business trips, said animals often die in the span of his absences. Eventually, when the need for another replacement occurred, Kress stumbled upon a terrarium filled with, what the shopkeeper described as, four colonies of *Sandkings*.

Each colony consisted of an immobile female, queen Maw, and a number of ant-like Mobiles, which are controlled by their Maw through telepathy. In order to survive, the tribe mothers need to be fed regularly, thus one of the Mobile's main purposes is to hunt down and collect food for her to digest. Luckily, the Maws are able to eat "anything" (other than sand), so everything that can be found in the terrarium can be brought to her. On the other hand, the Mobiles are not able to feed on their own, they can be only fed by the Maw, hence the second priority for mobiles is Maw defense, as when their mother dies, they will also perish. Throughout this paper we will use terms Sandkings and Mobiles interchangeably.

The shopkeeper informed Kress that, over time, the four colonies would start to wage wars between each other. Excited with this vision, Simon bought all four Sandking colonies and decided to have them installed in the living room of his flat. With time, the new owner hosted parties to show off his new pupils, and he couldn't wait for the conflicts to emerge. As Sandkings lived peacefully for the days to come, he started to starve them so they would become desperate.

From graceful and highly intelligent entities, the Sandkings turned into wild and murderous creatures that sought only to find more food to grow. Eventually, due to an unlucky event, they have broken out of the terrarium they were imprisoned in, and proved to be a threat not only to animals thrown into the terrarium, but also to their owner and to other people.

In our opinion, a story like the one which has been depicted above is not plausible, especially in the phase just before the Sandkings got free. However, being strangely attracted to the novel, we have decided to model the stage of the story, that took place slightly after Sandkings were in an extreme starvation period. This means that the only goal, which was driving the Sandkings, was to collect as much food as possible. In order to do so, the Mobiles not only would try to kill any living animals that were thrown into the terrarium, but would also fight with each other and potentially kill and eat other Maws – if that would prove profitable.

The needed model (and simulation) could most naturally be done using software agents. The application simulating the, defined above, situation could easily be used to predict plausible Sandkings behaviour(s). The main goal of our investigation was to resolve any doubts regarding lack of cannibalistic actions taken by the Maws towards "other tribes".

We proceed as follows. First, we shortly describe related work and tools used in our simulation. We follow by a description of the structure of the program and entities used in our simulation. Next, in Section IV, we list technological solutions used to develop the model. Following Section V, describes two scenarios and datasets used in them, (i) the friendly environment data set, and (ii) the deadly environment dataset. We conclude this section with overview of experimental results. Finally, we summarise our results present the resolution of our hypothesis.

A. State of the art

In nature, ants are social insects that, as individuals, are not capable of performing complex task(s), as they are "bounded" by their limited memory and behaviour that seems to have a noticeable random component. However, when the collective behaviour of ants is considered, maintaining pheromone-based communication, they prove capable of performing complicated tasks, like colony protection, or transporting food too heavy for an individual. This shows that even with very limited amount of computational resources per "agent" (ant), a swarm of ants can efficiently solve advanced tasks. Because of this,

researchers often make use of ant-like agent modelling, most often for optimisation purposes.

For instance, ant-like agents, used for load balancing in telecommunications networks, have been proposed by Ruud Schoonderwoerd, Owen Holland and Janet Bruten. If there is too much traffic through some network's node it might lead to loss of calls. Agents, whose behaviour has been modelled based on ants communicating by means of pheromones, traverse the network picking their path accordingly to pheromone distribution at each node, leaving an appropriate trail with each move. Such ant based model proved to be more efficient than shortest-path algorithms, or algorithm-based mobile agents [6].

Other use of ant-like agents is proposed by Stephen C. Pratt, David J. T. Sumpter, Eamonn B. Mallon and Nigel R. Franks in "An agent-based model of collective nest choice by the ant Temnothorax albipennis". The ants mentioned there are known for their ability to collectively choose the best nest site, out of several available, even if most of the ants have not visited more than one location. After finding a new site, the finding it ant tries to assess quality of the nest and convince the other active ants to move the colony. When new ants visit a location, adding to its temporary population, they reassess it by again trying to convince new ants. This might prove to be a kind of collective decision, where the best potential sites have the biggest temporary population. With proper empirical data, the algorithmic form of a collective decision-making mechanism can be captured and easily modelled [7].

Finally, the Biomass tool [8] allows to design ecosystem experiments. This agent based model simulation focuses on exploring the relationships between population in a given ecosystem. The individual decisions are based on environmental conditions. Finally the tool provides a way to configure the population by parametrization without having to program it manually. This work is the closest to our approach.

II. Tools

The *Repast Simphony* [3] is an open source agent-based modeling toolkit that simplifies model creation and experimentation. Out of a variety of accessible tools, we have chosen Repast for the development of the simulation. This was mainly due to its simplicity of use and the possibility of run-time dynamic interactions with the simulation. A big additional bonus for the choice of this modeling toolkit was the continued development and support of this package, which when compared to other agent-based modeling and simulation toolkits is a rare commodity.

Even though the *Repast Toolkit* supports many programming languages, including, among others, C#, VB.NET, Python, C++, Prolog, we decided to implement the application in the Java Runtime Environment. The most relevant arguments behind this decision include multi-operating system support, which Java provides and the fact that other programming languages were discouraged [5]. Apart from this, the developers of the *Repast Toolkit* suggest the usage of the *Eclipse* IDE, and thus we have followed their advice.

The creation process of 2D Euclidan environments and agents, with specific graphical properties, is particularly easy in the *Repast Simphony*. Specifically, an entity extending the class Agent is already ready for further development. Hence, the developer does not have to worry about implementation of the agent itself, there is only need to focus on general architecture of the agent based simulation.

In addition to this, another useful feature is a fully concurrent, discrete event scheduler, which provides a straightforward method of determining each agent type's behaviour on every simulation step.

Tutorial materials [2] provided by the Repast Team, especially the implementation of a zombie epidemic simulation, were the initial foothold in the topic of multi-agent based simulation. The extent and quality of document commentary significantly sped-up the process of simulation development.

III. Program structure

Agent-based computer systems are inherently object oriented, which allows for modularity and ease of development. In order for the simulation setup to closely resemble the universe created by *George Martin*, the modeling application has been divided into the following entities: Maw, Mobile, Formation, Food and Enemy.

A. Agents

Let us start with the description of the agent-based model, often referred to as ABM. Formally, it is a computational method that enables to create, analyse, and experiment with models composed of agents that interact within an environment. Most typically, they are used in social sciences, where researchers try to depict simplified versions of phenomena that are looked into. Based on a specific set of inputs, the model calculates resulting output data, which often are used to confirm or decline some theory [6], [8]. The implemented ABM is a system composed of multiple agents, which sustain interaction and communication. Such construct is called Multi-Agent system, otherwise referred to as a MAS. The agents relevant to our work are as follows.

1) Maw: The *Maw* (referred to as Mother or Queen) is an immobile agent, which spawns the Mobiles, described in the next section. Maw's life depends on the Mobiles ability to feed it, as once a Maw dies, so do all her Mobiles.

Four Mothers exist in our simulation (as shown in Figure 1: denoted as 1a, 1b, 1c and 1d). Every one of them, as it is supposed to exhibit a Hive-Mind behaviour, is responsible for coordinating the actions of her Mobiles. The children of each Maw inform her about every object found in the environment, possible threats and available food, allowing her to collect an extensive knowledge database. The types of threats and food are described in the Section III-B.

Based on the available information, each Maw organises the work of her children. This is done by means of a list of tasks (or assignments), which is sorted by the most profitable and least dangerous assignments being placed first. Such tasks have the purpose of keeping her fed, safe and alive. Possible

tasks are not limited to picking up heavy food, which single Mobiles are not able to move, or killing off monsters that can be later eaten, but also attacking other Maws in hopes for the acquisition of their food.

As picking up food is not related to any specific danger, except any possible threats met on the way, it is always considered the least dangerous and thus has the best profit potential. Attacking and killing an enemy is usually considered mediocre dangerous, where the profit is equal to amount of food that can be collected from the enemy and the danger is relative to the strength of the enemy. Finally, the profit of killing a Maw is considered. This involves the amount of food that can be gained by killing each of her children and the food that she actually has, while the danger is related to the overall strength of all her mobiles.

Depending on the strength, which is required to finish the task, the Maw might try to create a temporary alliance with another tribe. The lifespan of this alliance usually does not exceed the task greatly, as the profitability of another alliance may be calculated at any later stage of the simulation.

When one of the Maw's population grows bigger than the rest combined, it becomes a threat to other Maws. This can lead to an alliance constructed between the weaker Maws in order to start a war against the stronger fraction. Such behaviour is supposed to keep the strength of all Maws in balance.

2) Mobiles: The second most important element of the simulation are the Mobiles, as they act as a set of actuators for their Maw.

With simple autonomy, they prioritize the completion of the task they were assigned over their own survival, Mobiles wander around the terrarium looking for "points of interest". Every piece of food which is light enough to be picked up by a single Mobile is picked up and carried back to the fraction's Maw. Food, which is too heavy, and Enemies which are too strong to fight, are added to the Knowledge Base and are scheduled to be told to the Maw on the next occasion (i.e. when the Mobile goes home next time).

If an Enemy (in our simulation this is, most likely, a dangerous animal – see III-A3) is spotted, Mobiles will charge towards it, provided it is weak enough, and try to kill it in order to bring its leftovers to their Maw as food (see III-B2), or flee if fight would prove to be too dangerous.

As Mobiles are unable to digest on their own, they have to bring the food firstly to the Maw and only then she can feed them. Because of this fact, unless they have task to complete on the other side of terrarium, they tend to stay nearby their mother, keeping her safe and themselves fed.

3) Enemy : The opponents of Maws and Mobiles are the Enemies, also referred to as Monsters or Creatures. They are agents of much higher strength and health than a Mobile. They move on the Map in random directions, but do not eat anything they find, because their only purpose in the simulation is to fight with the Mobiles.

As shown in Figure 1, the following three types of Enemies exist in the simulation: *Spiders* [3c], *Scorpions* [3b], and the strongest ones – *Snakes* [3a]. As the simulation proceeds, the "damage level" of Enemies increases, depending on the current tick count of the simulation. The visual size of the enemies reflects their damage level (the more damaged they are, the smaller they get). This also indicates the number of Mobiles needed to kill them.

4) God: This is an "invisible agent" which has no graphical representation in the simulation. In each step, *God* may drop either Food, or an Enemy, in the terrarium, at a random point on the grid. The God "appears" every 100th tick of a simulation. Specific types of Food and Enemies to be spawned are chosen at random, where the probability of dropping them depends directly on their properties – the more powerful they are / the more food-value they have, the smaller the chance of their appearance. For example, *Pizza*, which is the most beneficial Food, for the Mobiles, and *Snake*, which is the strongest creature in the simulation have the least likelihood of them being spawned. Detailed properties of each entity, used in actual simulations, can be found in Section V.

The God module is also responsible for increasing the damage level of the enemies, as described in more detail in the Section III-A3. Additionally, a constraint was introduced such that the Monsters may never be created nearby the Maws, to prevent mobiles from dying immediately.

As opposed to the *Mobiles* and the *Maws*, neither *God* nor an *Enemy* have an ability to learn about the environment.

B. Environment

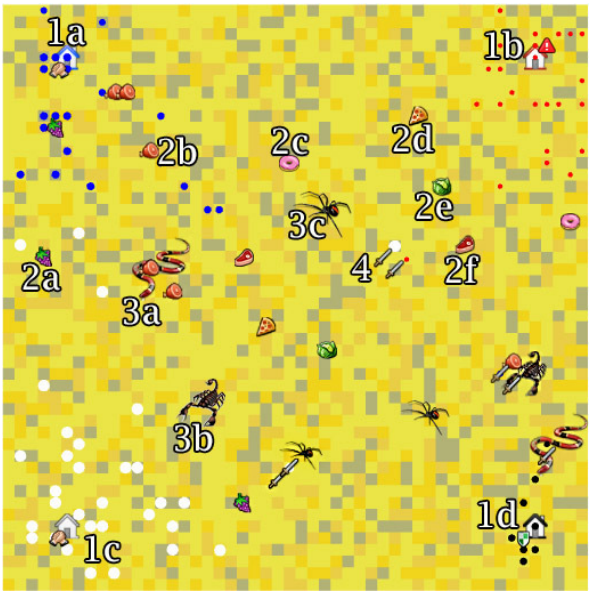


Fig. 1. Simulation with all elements.

The Environment of our simulation has been developed to represent the terrarium, from the *Sandkings* novel, as much as possible. As the simulation starts a Map representing sand and rocks is created, as well as Maws, which spawn their Mobiles during the runtime.

1) Map: The scenery of our simulation is a grid consisting of cells painted with different shades of yellow and grey, to graphically represent a simplified version of the terrarium, which was filled with sand and rocks. To make observation easier there exist icons representing an event shown for a given timeout (usually a few ticks – 50, 100, ... to grab attention but not to disrupt the visual flow of the simulation).

As seen in the Figure 1, the following icons – *Shield*[1d], *Shaking hands*[1a, 1c] and *Warning sign*[1b] concern Maws, hence they will appear right next to them. These icons indicate, respectively, that the Mother needs to be defended, has made an Alliance with another race, or that she is starving. The shield shows up when a Maw is in danger, which means that an opponent, either an Enemy or another fraction is about to attack her. In the case of the fourth sign, the *Sword*(Figure 1, [4]) will pop up next to every Mobile that is in a fight. The last informative picture is a *Grave stone* which appears on the Map in place of a dead Maw. Differently from the other icons, it does not have a timeout. At the end of a simulation, when only one Maw is left – as the others have died by hunger or during a war, an ending screen appears in the middle of a Map. It also contains information about which Maw was the last to stay alive.

2) Food : Food is an object (on the Map), which does not move. In our simulation, six types of Food with different, yet proportional, weight and "calorie count" properties are introduced. Weight property of each Food category affect how many Mobiles are needed to carry that piece of Food, while calories represent how much the "Maw's food repository" is increased by eating it. Using eaten food the Maw may either give birth to a new mobile or increase the overall tribe strength, making each Mobile stronger. However, note that the overall Food needed to feed the Mobiles may be larger thus introducing risk of starvation.

The first four types of Food are dropped by the God agent. As shown in Figure 1 they are represented with icons of *Grape*[2a], *Doughnut*[2c], *Pizza* [2d] and *Cabbage* [2e] .

The last two types of Food are placed in the grid only (and every time) when either a Maw, Mobile or Enemy dies. When a Mobile dies, it turns into one piece of Food, shown as *Meat* (Figure 1, [2b]). For balancing purposes, a piece of Meat has less calorie value than any other kind of Food, and amounts to roughly half of the upkeep needed to feed one Mobile. Both Maws and Enemies become *Steaks* (Figure 1, [2f]) after their death. This means that the strength of an agent is directly proportional to the Food this agent drops when dying.

Additionally, dead Maws also drop extra Steaks proportionally to the amount food she has eaten throughout the entire simulation. The detailed values of each Food's properties, used in actual simulations, has been specified in Section V.

C. Formations

In order to model agent cooperation, a simple Mobile formation logic has been implemented. Specifically, a temporary agent called Formation is spawned for the duration of a specific Task. Every Mobile, which is in a Formation, is not allowed to carry out any of the movement, item carrying or fighting logic associated with a single Mobile. Instead, the Formation object is responsible for the simulation algorithms of all Mobiles in its ranks.

A new instance of a Formation gets instantiated when one of the Maws has a Task, in her Scheduler, which is too "difficult" for a single Mobile, i.e. *object too heavy to carry* or *enemy too dangerous to attack*. This allows for easy planning, basic tactics and task assignment. Formations are assembled by the respective Maws and are provided with all relevant information (during assembly of the Formation): the Task type (fetch food, attack an enemy), the size needed for the Formation to be formed and the Task location (the last known position of the food or enemy). To find the number of Mobiles necessary for the creation of a Formation, the Maw looks for any of her children, which are currently not assigned to any other Task, and are in her vicinity.

If the Task needs more Mobiles than available to the Maw (at a given moment), the Alliance system comes into play – Mothers of different fractions are asked whether they want to join the Task. If the asked Maw deems this profitable, she creates a Formation of her own. When the critical number of Mobiles for the given Task is reached, the Formation assembly process is finished and the Formation is sent out to carry out its Task. All Formation in a multi fraction Alliance have their movement synchronised so that they will arrive at their destination at the same time.

A Formation is created "on the position of the Maw", which it belongs to. This is due to the fact that she has to have the possibility to oversee the Formation assembly process by assigning free Mobiles from its vicinity to the Task. The Formation waits until the Maw finds enough Mobiles from either her own ranks or through Alliances and all assigned Mobiles arrive at the location of the Formation.

A Formation, which has not yet arrived at its goal location, moves all its assigned Mobiles towards the goal location (in this case, the food or enemy, which was supposed to be picked up or attacked). If the Formation's goal is attacking something, all profitable Enemies in the vicinity are attacked in addition to moving to the set goal position. Similarly to the logic of the Mobiles, a Formation can have specific functions called on arrival at a specific location. For example, this could check the Food in the vicinity, pick it up and start heading back for home. In another example, all Mobiles could attack an Enemy, provided one is nearby.

IV. Methodology

In order for an agent to act with a logic driven approach, a concrete set of "tools/methods" has to be introduced. Here, tools/methods, used in our simulation, are described.

A. Fighting

Relations between each tribe in terrarium can be described as Hostile, Neutral or Friendly, with tribes dividing into each Maws faction and Enemy tribe. Initially, the relation between each of mentioned tribes is set to Neutral.

When a mobile stumbles upon other Mobile, Maw or Enemy, if the relation between them is not friendly, a fight between them will start. If an agent is fighting, instead of moving, during the tick it inflicts damage equal to its attack parameter to one nearby non-friendly agents, i.e. reduces its amount of health by the value of attack, prioritizing hostile agents. When there are no non-friendly agents left in the vicinity, fight ends with a victory. If agents nearby are too dangerous to fight, it flees. In case of victory, the Mobile restarts its previous task, in case of fleeing it will move in the other direction than the danger is in, as to avoid it.

When mobiles are grouped into a Formation, as described in Section III-C, they are not executing their individual tick steps. Instead, the Formation is managing their actions, acting accordingly to its own tick step(s). Similarly to mobile fighting, the fight starts if Formation encounters non-friendly agents nearby and might end either in victory if there are no non-friendly agents left, or a retreat decision. Formation decides to retreat from fight if there are less mobiles left than the half of initial Formation size. During the fight instead of moving during its tick step the Formation orders each subordinate mobile to attack, as it was itself in a fight.

B. Message Exchange

We have constructed a simple mechanism to allow for an inter-agent communication. This system does not use the communication functionality provided by the *Repast Simphony*, as we had to create a "non-standard communication mechanisms" (to match the need of our simulation).

Agents can send messages to each other. This is done by adding an abstract message object (i.e. packets – which have sender and recipient) to a global message queue. This queue, on notification, will process its contents, and fire specific handlers for each message type. Based on the sender's needs and the recipient's current state, a response is sent (or not).

In our simulation, Mobiles and Maws communicate between each other on regular basis – as often as the need for communication arises (for example, when a Mobile has found knowledge or is hungry). Since each Maw is an immobile agent, all knowledge about the environment comes from its children. In order for a Mobile to send any message, it has to be in a small distance from its potential recipient. Thus, if a Mobile wants to message its Maw, it has to "return back home". Upon return, a Mobile can send a specific message to its Maw. Sending an *Inform* message will simply cause the Maw to add a given piece of information to its Knowledge Base. Mobile can also send an *Ask For Food* message and wait for a response. In this case Maw can respond with either *Acceptance* or *Rejection* based on its food supplies.

Being true to the book, Maws have the ability to send messages without any distance restrictions (through telepathy). A given Maw can also start communicating with other Maws, if it is faced with a problem it can not solve alone, e.g. dealing with a strong Enemy creature. In such case, a series of *Ask For Alliance* messages is exchanged between all Maws in order to begin an Alliance and to send out Formations, which could potentially defeat this Enemy.

C. Knowledge Base

Both Mobiles and Maws gather information about their environment. A single piece of information encapsulates the following properties: the object of interest, its location and the time when it was discovered. This information is saved in, unique for every Agent, Knowledge Base.

Agents do not duplicate information in their knowledge bases. Upon encountering an object of interest, the Mobile checks whether it already knows about it. Each piece of information in the knowledge base can be marked as *useless*. Useless information is the one that has already been processed and no further action should ever occur based on it. This is explained in detail in section IV-D.

Mobiles explore the Map and learn about their surroundings. Each time a Mobile encounters something interesting, an information about this object is added to its knowledge base. Each Mobile can learn about other hostile Mobiles, Enemy creatures or Food scattered around the Map. Once they find something, they run back to their Maw to inform her about their findings. Both Maw and her Mobiles have their own Knowledge Bases.

As mentioned previously, Maws can not explore on their own. Thus all information comes from the Mobiles exploring the Map. The Knowledge Base is used to control the actions of the Maw's children; see, Section IV-D.

D. Scheduler

The Agents' behaviour is scheduled dynamically during the simulation runtime. To achieve this, each Agent is given its own Scheduler, responsible for assigning Tasks, based on its Knowledge Base. It runs in parallel with their Knowledge Base by going through *non-useless* information and determining the next course of action.

A Task is a sequence of steps, which an agent should follow in order to complete a task. Tasks are split into stages. A stage is built of an execution part, which determines the necessary steps of the agent, and the condition, which indicates the end of this stage, thus indicating the beginning of another one. When the task reaches its final stage, it finishes successfully, which in turn marks the information that initiated this task as *useless*. A Task can also be finished successfully in abnormal cases, when a given stage was impossible to execute. Abnormality occurs, for instance, when an Agent reaches the destination and the object of interest is no longer there. It could mean that similar Task has been completed by different Agent. However, the current task can be replaced with another one having higher priority. In this case, the current Task is stopped, but the information is kept in the Knowledge Base without modification. This allows to restart the Task when the Scheduler, in synchronisation with the Knowledge Base, has determined that the Task should, nevertheless, be completed.

For example, a Mobile can be assigned a Task to pick up Food from a location, which is stored in its Knowledge Base.

This task is split into the following stages: move to Food's location, pick up the Food, and return home. When the Mobile returns home safely, the task is finished successfully. Since Mobiles do not share any global knowledge, other Mobiles of the same fraction could already haven returned this Food. Of course, the same applies to Mobiles between different fractions. Hence it can happen, that the indicated location contains no Food at all. In this case the $pickupfood$ stage fails and this task is marked as finished successfully. Even though the Mobile did not deliver the food we do not want the Mobile to repeat the process of delivery when there is no food to be picked up.

A Mobile can also be assigned a Task to inform its Maw about given object. This can occur when the found Food is too heavy for it to pick it up alone, or when it sees an Enemy Creature or Formation. During execution of this Task, a Mobile will try to avoid combat, if only possible. Of course, it may happen that the returning Mobile encounters an Enemy and gets killed by it. In such case the information it "carried" is forgotten and has to be rediscovered by other Mobiles.

After being informed about the environment, Maw can generate a Task, which will control Mobile units. Execution of tasks for picking up heavy Food and fighting Enemy creature(s) has similar structure. In the first stage, Maw has to determine if given objective can be completed. In other words, if it has enough Mobiles to pick up food or to defeat certain Enemy. Then a Formation is constructed, which starts moving towards its objective. In case of Food, the formation is ordered to pick it up and return home. When the objective is to defeat an Enemy, the formation has to destroy it. In both cases, at the end of task the Formation is disbanded, and Mobiles return to normal work that is to explore the map.

A Maw can also start a War Task. As defined previously, War can occur when one of the fractions becomes stronger than all other fractions combined. When this is true, a single Maw can start communicating with other Maws using the *Ask For Alliance* message in order to create a force that could weaken the common Enemy Maw.

When a Maw is informed about an Enemy Formation, a Defend Task is initiated. This will cause all Mobiles to return home immediately to form a Formation of their own in order to destroy their (incoming) enemy.

V. Simulation

Now, let us describe the tested input values and the results of the preliminary simulations. The data set of our application was divided by harshness of the environment into two scenarios: *friendly* and *deadly*. Based on this division, two simulation sets were conducted, each returning different results. The comparison of said results leads to interesting observations regarding our main thesis.

In our opinion (hypothesis we started with, and tested in our simulations), cannibalism is a natural outcome of profit-driven logic, in which the only relevant factor in the decision making process is the amount of food that can be gathered as a profit, as it vastly increases the survival chances of an individual Maw.

This, however is not presented in the *Sandkings* story, where the survival of the race often had a higher priority than the survival of the individual. This leads to our hypothesis that with progressively decreasing friendliness of the environment, the aggression level towards Mobiles of other fractions, and other Maws, in particular, vastly increases.

In a friendly environment, Maws usually do not need to rely on Alliances, as they are strong enough, on their own, to cope with most threats in the terrarium. When the environment is much stronger than their respective power levels, cooperation – even if temporary – is the Maw's only chance of survival.

When the terrarium is filled with enough food and relatively weak Enemies, an endless simulation, i.e. prolonged lifespan of each race without clear victory of one of them – should be possible, as shifts in power balance between the fractions should be much smaller than within a deadly environment. This means that, most probably, with harsh configuration, a victorious Maw will quickly emerge.

A. Shared Data set

In the simulation we have used four Maws with their respective Mobiles, six different Food types, three Enemy types, and a simplified automatic spawning algorithm called the *Autogod*. The grid size was set to 50x50. This structure allowed us a simplified, yet visually engaging, representation of the environment described in the original novel.

Enemies' graphical size changed with time, accordingly to their health and attack level increased with time. The first was increased by the 4th square root of the tick count and the latter – by the 5th. There was 10% chance that an Enemy will be dropped on each God's step. In Tables I, II, III we present the input data sets that are constant in both Friendly (Section V-B) and Deadly (Section V-C) environments.

TABLE I
MAW PROPERTIES

Property	Value	Comments
strength	0	Strength at the beginning of the simulation.
food unit	0	How much food calories a Maw has at the start.
attack	0	Initial number. Changes depending on Maw's strength.
health	1000	Initial number. Changes depending on Maw's strength.
meat count	50	How much steak is dropped when it dies.

TABLE II
MOBILE PROPERTIES

Property	Value	Comments
steps per food	150	How much steps it can take until it starves.
stomach size	2	How much calorie at a time it can eat.
attack	5	How much damage it makes per one attack.
health	100	How much damage it can take before it dies.
meat count	1	How much meat is dropped when it dies.

TABLE III
FOOD PROPERTIES

Property	Weight	Calorie
Cabbage	1	10
Grape	3	30
Doughnut	5	50
Pizza	7	100
Meat	1	3
Stake	1	20

As discussed above, the Weight determines how many mobiles are needed to carry each food. Both properties – weight and calorie – are constant during the span of a simulation.

B. Friendly Environment Data Set

In the first simulation we have defined what we believed to be a friendly environment; one in which it is easy for the Mobiles to survive. Here, in the *Autogod* module, a 33% of chance is given that a food will be dropped on each of God's step. While the probability of spawning Enemies is equal in both environment types, in the friendly environment their strength and health is optimized so that the Mobiles should not have difficulty in eliminating them (see, Table IV).

TABLE IV
FRIENDLY DATA SET: ENEMY PROPERTIES

Enemy Type	Spider	Scorpion	Snake
Attack	10	20	40
Health	150	300	600
Meat value	1	4	20

C. Deadly Environment Data Set

This environment was set so that it would be hard for the Mobiles to survive. The purpose of this experiment was to observe how long they can live under harsh circumstances. In addition to this, the occurrence of wars and alliances in comparison to the Friendly Environment, was expected to change. The probability of Food being spawned by the Autogod has been dropped to 10%, which makes surviving much harder for the Sandkings. The Enemies' properties were also adjusted according to our vision of the harsh environment (see, Table V).

TABLE V
DEADLY DATA SET: ENEMY PROPERTIES

Enemy Type	Spider	Scorpion	Snake
Attack	15	30	60
Health	200	450	800
Meat number	1	4	20

VI. EXPERIMENTAL RESULTS

The following tables describe the results of 20 sample simulations (10 in friendly, 10 in harsh environment settings). Average values of each of the simulation types are presented. Note that we have not observed large variation in results. Henceforth, average values presented below are "representative" to both sets of experiments.

A. Simulation with the Friendly Data Set

As shown in Table VI, the application of the Friendly dataset (see Section V-B) resulted in simulations ending, on average, after 7971 ticks, spawning almost 40 Enemies and dropping 134.6 Food pieces on the map. The Maws have formed, on average, 3.2 alliances per simulation.

TABLE VI
FRIENDLY SIMULATION - GENERAL STATISTICS

GENERAL	Tick	Enemies spawned	Food dropped	Alliances
Average	7971.6	39.9	134.6	3.2
Per Tick	−−	0.005	0.017	0.00040

In addition to this, the statistics about the average, winner (last survivor) and losers are presented in Table VII. The columns describe the time when the death of one of the Fractions occurred, the number of lost Mobiles and the peak value of the Mobile count, and the amounts of Food consumed.

Finally, a third statistic concerning the number of Tasks (and by this, Formations) is shown in Table VIII.

TABLE VII
FRIENDLY SIMULATION - MAW STATISTICS

MAW	Death Tick	Lost	Max	Meat	Non-meat
Total AVG	2639.20	76.53	42.33	116.55	23.30
Per Tick		0.01	0.01	0.01	0.00
Winner AVG		89.13	51.25	136.88	30.63
Losers AVG	3650.56	73.52	38.11	103.11	19.59

TABLE VIII
FRIENDLY SIMULATION - TASK STATISTICS

TASK	Food	Creature	Defend	War
Total AVG	23.00	16.93	0.10	0.25
Per Tick	0.00	0.00	0.00	0.00
Winner AVG	41.75	27.88	0.13	0.50
Losers AVG	16.63	13.74	0.11	0.15

B. Simulations with the Deadly Data Set

In the Deadly Data Set – as shown in Table IX – the simulation ended, on average, after 3599 ticks, almost equally spawning 18 and 18.6 Enemies and Food pieces on the map. The Maws have formed 4.5 alliances per simulation.

This simulation run shorter than the friendly one, which means that the Mobiles faced a bigger difficulty level. The comparison of the alliance count also shows that, under these circumstances, the Maws are more likely to form an alliance to be able to combat a common enemy. Taking into account the difference between length of simulations and number of alliances made in Friendly and Deadly environments, in the first case alliances were made approximately every 2491 tick count and in the latter - every 800. This also means that when

facing harsh simulation, the Maws formed over 3 times more alliances.

TABLE IX
DEADLY SIMULATION - GENERAL STATISTICS

GENERAL	Tick	Enemies spawned	Food dropped	Alliances
Average	3599	18	18.6	4.5
Per Tick	−−	0.005	0.005	0.0013

The Maw statistics (see Table X) again show the time of death, lost and peak amount of Mobiles and the eating habits of the overall, winner (last standing) and losers. The lifespan per Fraction does not differ significantly from the Friendly Data Set, but the amounts of Mobiles and Food consumption are proportional to the harshness of the environment and resulting duration of simulation.

TABLE X
DEADLY SIMULATION - MAW STATISTICS

MAW	Death Tick	Lost	Max	Meat	Non-meat
Total AVG	2074.25	34.20	17.73	40.95	3.75
Per Tick		0.01	0.00	0.01	0.00
Winner AVG		26.50	26.00	53.50	5.20
Losers AVG	2765.67	36.77	14.97	36.77	3.27

In direct comparison, the Task statistics (as presented in Table XI) differ significantly. The Maws, when placed in a Friendly Environment, sent more of their offspring to fetch Food or to attack nearby Enemies, though this is most probably due to the scarcity of the Food and the strength of the Enemies.

TABLE XI
DEADLY SIMULATION - TASK STATISTICS

TASK	Food	Creature	Defend	War
Total AVG	3.10	5.13	0.13	0.25
Per Tick	0.00	0.00	0.00	0.00
Winner AVG	4.20	7.10	0.00	0.30
Losers AVG	2.73	4.47	0.17	0.23

The Deadly Data Set made the Maws focus their attacks *not* on the Creatures dropped into the Terrarium, but rather on the Mobiles of other Fractions as they were weaker and easier to eat. In addition, the winning Maw is more likely to have started a war – though this is only visible in the Friendly Simulation.

In the case of the Friendly Data Set, Maws died mainly due to starvation, where the winning Fraction would collect more than the other three. The Deadly Data Set showed some examples of Enemies (or, enemy Formations) attacking one of the Maws directly, leaving the defenders not enough time to prepare – leading to substantial losses or even death of the Sandkings tribes.

Most of the time, the first Maw, with the weakest start, died rather quickly in the simulation (as described in the book). The second Fraction to fall usually lasted until approximately 3/4 of the simulation, leaving the surviving two living mostly peacefully. In rare cases, observed in multiple simulations run during program development, all four died in a few hundred steps.

Alliances of three Maws against the fourth occurred very rarely, proportionally to the same strength growth for all Fractions. The Alliance system was most often used to attack Creatures in the Terrarium. A stronger Maw attacking one of the weaker ones, in order to gain Food, only happened rarely – when a Formation, which was sent out to attack a Creature, finished its task and started chasing another Fraction's Mobiles running into their Maw and killing her.

VII. CONCLUSIONS

Analysis of results from both Friendly and Deadly Data Sets led to surprising results. Our initial hypothesis was that, with the increased hostility of the environment, the Sandkings would develop cannibalistic behaviours and fight until eventually only one tribe is left.

Contrary to it, with harsher environment the Sandkings showed tendencies to cooperate in order to increase their own chances of survival, and refrained from attacking each other, as the sustained losses most probably would prove incomparable to the potential food gain. On the other hand in friendlier environment, where Sandkings amassed sizable amounts of food and grew in size and numbers, raiding other Maws proved profitable enough that acts of cannibalism have been observed.

Concluding, our hypothesis was faulty, as our Sandkings model proved that their aggressiveness increases not with harshness of environment they are located in, but with its friendliness. With the results presented, it is safe to state that the vision presented in work of George R. R. Martin was plausible, as Sandkings when facing extremely unfriendly conditions had to cooperate in order to survive.

ACKNOWLEDGEMENT

We would like to thank Marcin Paprzycki and Maria Ganzha for a fascinating introduction into this topic, wise guidance and constant kindness.

REFERENCES

[1] Martin, G. (1981). Sandkings. New York: Pocket Books.
[2] Collier, N. and North, M. (2015). Repast Java Getting Started. 1st ed. [ebook] Repast Developement Team. Available at: http://repast.sourceforge.net/docs/RepastJavaGettingStarted.pdf [Accessed 30.05.2015].
[3] Repast.sourceforge.net, (2015). Repast Suite. [online] Available at: http://repast.sourceforge.net/ [Accessed 30.05.2015].
[4] Railsback, Steven F., Steven L. Lytinen, and Stephen K. Jackson. 'Agent-Based Simulation Platforms: Review And Development Recommendations'.
[5] Crooks, Andrew. 'An Introduction To The Repast Software Recursive Porous Agent Simulation Toolkit'. 2015. Presentation.
[6] Schoonderwoerd, R., Holland, O., Bruten, J. and Rothkrantz, L. (1997). Ant-Based Load Balancing in Telecommunications Networks. Adaptive Behavior, 5(2), pp.169-207.
[7] Pratt, S., Sumpter, D., Mallon, E. and Franks, N. (2005). An agent-based model of collective nest choice by the ant Temnothorax albipennis. Animal Behaviour, 70(5), pp.1023-1036.
[8] Candelaria E. Sansores, Flavio Reyes, Hector F. Gómez, Juan Pavón, Luis E. CalderÃłn-Aguilera. BioMASS: a Biological Multi-Agent Simulation System. Proceedings of the Federated Conference on Computer Science and Information Systems pp. 675–682.

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

THE EMERGING smart infrastructure in energy networks represents a major paradigm shift in resource allocation management with the aim to extend the centralised supply management model, towards a decentralised supply-and-demand management that is expected to enable more efficient, reliable and environment-friendly utilisation of primary energy resources.

Together with this vision, there are new and complex tasks to manage, in order to ensure safe, cost-reducing and reliable energy network operations. This includes the integration of various renewable energy systems, like the photovoltaic or the wind energy, which are able to reduce the greenhouse gas emissions but that are working under greater uncertainty; as well as the interaction of transport and storage systems for energy that are envisioned through techniques like 'Power to Gas' and fuel cells, which are using the electrical and the gas transportation network.

Further tasks can be found in the fact that the market participants (e.g. simply households) are becoming more autonomous and intelligent through technologies like smart metering, which requires a coordinated demand side management for millions of producers, consumers or, if this applies, prosumers by means negotiations and agreements.

Information and communication technologies are key enablers of the envisioned efficiencies, both on the demand and the supply sides of the smart energy networks, where the agent-paradigm provides an excellent first modelling approach for the distributed characteristic in energy supply systems. On the demand side they aim at supporting end-users in optimising their individual energy consumption, e.g. through the deployment of smart meters providing real-time usage and cost of the energy and the use of demand-response appliances that can be controlled according to the user preferences, energy cost and carbon footprint. On the supply side they aim at optimising the network load and reliability of the energy provision, e.g. through active monitoring and prediction of the energy usage patterns, and proactive control and management of the reliable energy delivery over the networks. It is also envisaged that they will be able to influence the demand through the dynamic adjustments of the energy price in order to influence the end user behaviour and energy usage patterns throughout and across the energy networks for electricity, gas and heat.

Although a significant effort and investment have been already allocated into the development of smart grids, there are still significant research challenges to be addressed before the promised efficiencies can be realised. This includes distributed, collaborative, autonomous and intelligent software solutions for simulation, monitoring, control and optimization of smart energy networks and interactions between them.

Topics

The SEN-MAS'15 Workshop aims at providing a forum for presenting and discussing recent advances and experiences in building and using multi-agent systems for modelling, simulation and management of smart energy networks. In particular, it includes (but is not limited to) the following topics of interest:

- Experiences of Smart Grid implementations by using MAS
- Applications of Smart Grid technologies
- Management of distributed generation and storage
- Islands Power Systems, Microgrid Applications
- Real time configurations of energy networks
- Distributed planning process for energy networks by using MAS
- Self-configuring or self-healing energy systems
- Load modelling and control with MAS
- Simulations of Smart Energy Networks
- Software Tools for Smart Energy Networks
- Energy Storage
- Electrical Vehicles
- Interactions and exchange between networks for electricity, gas and heat
- Stability in Energy Networks
- Distributed Optimization in Energy Networks

Event Chairs

Derksen, Christian, University Duisburg-Essen, Germany

Kowalczyk, Ryszard, Swinburne University of Technology, Melbourne, Victoria, Australia

Steering Committee

Derksen, Christian, University Duisburg-Essen, Germany

Lehnhoff, Sebastian, OFFIS - Institute for Information Technology, Germany

Kowalczyk, Ryszard, Swinburne University of Technology, Melbourne, Victoria, Australia

Nahorski, Zbigniew, Systems Research Institute - Polish Academy of Science, Poland

Program Committee

Andrew, Lachlan

Braubach, Lars, University of Hamburg, Germany

Guttmann, Christian, Monash University, Australia

Klusch, Matthias, German Research Center for Artificial Intelligence, DFKI, Germany

Linnenberg, Tobias, Helmut Schmidt University, Germany

Moench, Lars, FernUniversität Hagen, Germany

Ossowski, Sascha, University Rey Juan Carlos, Spain

Sonnenschein, Michael, Carl von Ossietzky University, Oldenburg, Germany

Sudeikat, Jan, Hamburg Energie GmbH, Germany
Unland, Rainer, Universität Duisburg-Essen, Germany
Weber, Christoph

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 259–264
DOI: 10.15439/2015F396
ACSIS, Vol. 6

Energy Agents - Foundation for Open Future Energy Grids

Christian Derksen
DAWIS - University of Duisburg-Essen, Schützenbahn 70, 45127 Essen, Germany
Email: christian.derksen@icb.uni-due.de

Rainer Unland
DAWIS - University of Duisburg-Essen, Schützenbahn 70, 45127 Essen, Germany
Email: rainer.unland@icb.uni-due.de

***Abstract*—Redefining our energy landscape by means of regenerative, volatile and decentralized organized systems represents a major challenge for the creation of distributed control solutions. Standards are required that permit describing the variety of energy conversion systems and that enable an interoperable exchange of energy amounts in an open, flexible manner by concurrently taking into account technical and market requirements. This paper introduces the concept of unifying Energy Agents. They have the potential to reduce the ever growing complexity that comes along with the various solutions presented or that are already available at the market. In contrast to that, the notion of Energy Agents stands in our view as a representative for a required methodology that enables a consistent development of de-centralized control solutions. In order to demonstrate this approach, the application of Energy Agents in a smart house scenario is discussed. It is shown how arbitrary agents with different levels of sophistication and abilities can cooperate with each other in a smooth way. It is our strong believe that a stepwise and standardized development of Energy Agents representing the needed decentralized control solutions is needed for a sustainable design of an open Future Energy Grid.**

I. Introduction

Reflecting the tendencies for more decentralized controlled energy conversion systems and the further increasing number of IT-enriched smart systems in general, a picture can be drawn, that describes the future energy landscape as a complex, globally connected and mainly software driven system. Smart Markets need to build on top of an underlying technical systems with inherent flexibility that guarantees a stable volatile energy production [1], in order to reach climate targets [2] or just to maximize organizational profit [3]. However, global goals, such as the stabilization of a distribution networks, require a minimum of adaptive interoperability that has to be expressed in one standard. We believe that the developments in Smart Grids and related areas are now at a point, were it has to be asked, if we want to build control systems that are creating new monopolies, caused by proprietary software solutions, or if we want to create an unbundled and open energy supply that, on the one hand, offers the needed intelligent flexibility and, on the other hand, strongly supports further developments over the next decades? Assuming that the latter is the case, it is obvious that software standards are required that prevent our with respect to technical basics and market regulations already highly complex energy supply from becoming even more complex and possibly uncontrollable.

With the concept of unified Energy Agents and its associated frameworks and tools that are a unifying Energy Option Model (EOM) [4] and the agent execution environment Agent.GUI [5], we provide a development and validation environment that allows a systematic and stepwise progression for decentralized control solutions. We believe that a stepwise definition of concrete use case scenarios and a clear definition of the actual tasks and capabilities of the on-site software - that are in our view Energy Agents - are strongly required. Corresponding to these different scenarios, Energy Agents have to be defined with specific levels of sophistication that have to be validated through simulations and in test-bed applications before they can be applied in real on-site systems. Furthermore, energy carrier like natural gas, heat and other have to be considered in order to define a comprehensive problem description that reflects the complex and interconnected nature of the real energy supply.

This paper presents an application example of our Energy Agent approach. For this, the next section provides the theoretical background. This will be followed by the presentation of the concept of Energy Agents. Section 4 presents their application in a smart house scenario. A discussion and an outlook will conclude the paper.

II. Background

Agents and Multi-Agent systems (MAS) are already used by a large scientific as well as industrial community due to their inherent ability to naturally describe the distributed problems and scenarios that comes along with the energy domain. A significant number of papers already described successful applications of agents in specific Smart Grid scenarios, as for example in virtual power plants [6], in Demand-Side-Management systems [7] or within price-based, indirect controlled approaches that are known as Demand Response [8]. Despite these promising first

applications connecting IT-concepts and standards are still missing that enable a large-scale rollout of Smart Grid solutions by concurrently providing a meaningful function and an investment security for end users and producers. The problem to be tackled here is a missing open architecture. Right now there is the danger of new monopolies, caused by investment decisions towards proprietary Smart Grid solutions that may preclude customers from switching to a new energy provider. Another threat of an uncontrolled growth of arbitrary (software) solutions is the resulting complexity and a possible in-deterministic behavior of the overall system, especially if every software component in a Future Energy Grid can behave as it likes.

Such types of problems have already been discussed in the scientific community that generally has been studying the building of agent organizations and the related topics of reputation and trust [9]. Following [10], MAS-organizations can roughly be categorized by two dimensions. The first dimension permits to differentiate between MAS that rely on existing parent organizational structures or do not. This classifies designs of MAS into so called "agent-centered" or "organization-centered" designs. The second dimension focusses on the abilities of agents. It distinguishes by whether they have organizational knowledge or not. Especially the latter dimension is to be equated with the question if an agent organization represents a concept that only serves the static design of a multi-agent system or if the organizational affiliation is relevant at the runtime of the individual agents. Table 1 below shows the resulting types of agent organizations derived from these two differentiations.

TABLE 1: CLASSIFICATION OF AGENT ORGANIZATIONS [11]

	Agents without organizational knowledge	Agents with organizational knowledge
Agent-centered Design (no given formal organizational structure)	**I.**	**II.**
Organization-centered design (given formal organizational structure)	**III.**	**IV.**

Without discussing every type of agent organization in detail here, we would like to point out that in quadrant IV. open agent organizations are located, where agents are able to dynamically decide if they want to join or leave an agent organization. To support functionalities like this, further efforts has to be spent in order to enable such dynamic and thus adaptive behavior. We believe that this is - to some extent - indispensable for the further systematic development of an open, IT-controlled Future Energy Grid. Consequently, this requires commitments and standards that focus on the local software components on-site that we propose to call Energy Agents.

III. ENERGY AGENTS

The concept of Energy Agents was first introduced by [12]. There an Energy Agent was defined as a representative of a technical system - or better an energy conversion process - that acts on an operational level and that will not replace but extend on-site controllers. Thus, Energy Agents can be seen as additional software artefacts, capable to autonomously manage the capacitive abilities of associated technical systems in terms of energy production and usage and on behalf of the owner or other stakeholders. According to the tasks assigned to it, the inherent complexity and thus the sophistication or so called Integration Level (IL) of an Energy Agent may differ; different tasks, like monitoring, learning or constraint satisfying planning and optimizations are conceivable in this context.

Furthermore, an approach to validate decentralized control systems was proposed by means of a systematic development process that includes simulations and test bed-applications before an on-site hard- and software usage should be considered. By implementing exchangeable behaviors for simulated or real world interactions between agents and local system controllers, the main software components of the Energy Agent are to be preserved in regard to the development process. Finally, a concept for a generalized option, cost and action model was introduced that is capable of describing possible and current operation phases of technical systems and that serves the Energy Agents as a unified base model for reasoning processes. In these considerations also hybrid energy systems and infrastructures were taken into account, as for example natural gas or heating systems.

Figure 1 below gives an impression and compares two types of Energy Agents, equipped with different capabilities. It is shown that the shell that represents the Energy Agent is coupled in two ways to the outside world. First, by means a connection to a technical system and second by it capabilities for inter-agent communication. Here the coupling to the technical system can be realized in various ways; e.g. by means of well-known protocols like *IEC61850*, the *Common Interface Modell* (*CIM*), *OPC UA* and other.

Following our proposed gradation of integration levels (see reference above), both Energy Agents have a domain specific model - the Energy Option Model (EOM) - that describes the operating capabilities and the scope for possible actions of the underlying technical system. Both agents can be located in IL3, as both have the ability to (re-) act on external signals or information in general. Beside that they have the ability to monitor the underlying technical system and provide information that help to predict the systems behavior in regard to its energy production or usage, which is recommended by the introduced lower integration level IL0 - IL2.

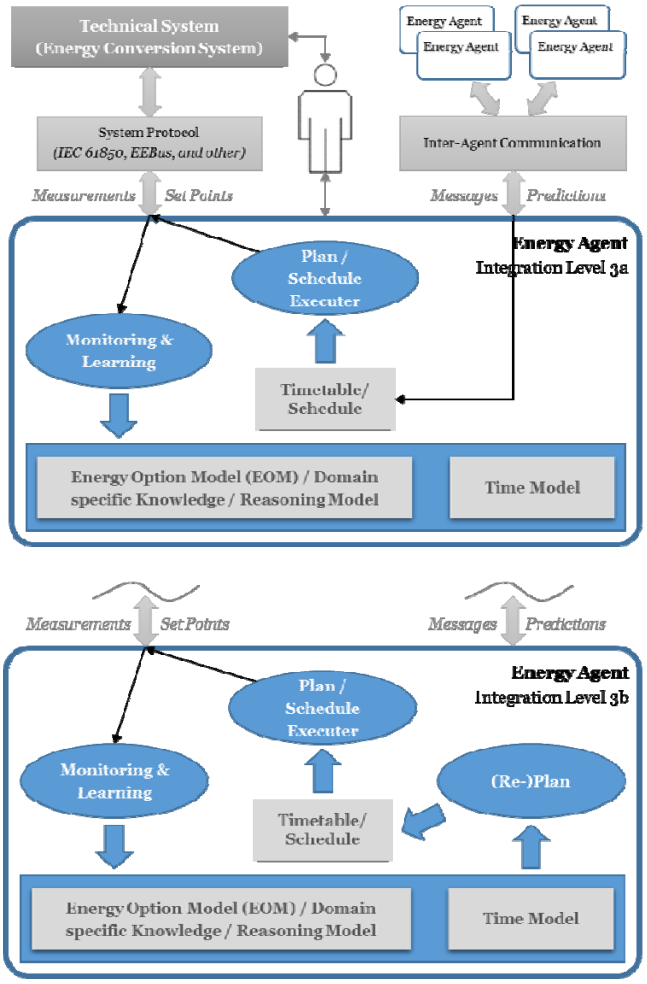


Fig.1: Structure and functionalities of Energy Agents with different Integration Level

Furthermore, the picture underlines that in regard to the requirements for an open Future Energy Grid the definition of the actual application scenario is absolutely indispensable, since the functionalities that are to be implemented within an Energy Agent strongly depends on these scenarios. Even both Energy Agents are located in IL3, the upper agent just receives and accepts schedule from a superordinate control unit, while the bottom Energy Agent uses the EOM in order to determine an optimized schedule by itself.

With the Energy Agents above, two different Smart Grid approaches are described that have often be used and discussed in scientific publications. While the first Energy Agent could be applied in a Demand Side Management scenario, where a central process decides for operational states and set-point configurations, the second can be used within a more decentralized Demand Response approach, e.g. reacting on external price signals.

It is crucial to realize that in both cases the Energy Option Model, purposed with the Energy Agent approach, forms the base for both application cases. With its approach that allows to uniformly describe energy conversion systems, the EOM allows to transfer and relocate decision making processes and thus to realize different types of control solutions; either centralized or decentralized solutions. We believe that such unified descriptions is the first step towards the realization of an open Future Energy Grid, since it allows an adaptive management of net-coupled technical systems or energy conversion processes. As a second step, the application scenario with its corresponding rules, policies and behaviors have to be standardized, so that developer of Smart Grid solutions have not only requirements but also degrees of freedom for their developments. Therefore, the integration levels proposed have to be fully specified, while the levels needs to be developed and extend over the next years.

IV. Application of Energy Agents

With the application scenarios presented here, we demonstrate different level of sophistication for Energy Agents by means of comparing Demand Side Management with a price-based Demand Response approach. As use case we selected a smart house, equipped with different technical systems that are each “interfaced” by an Energy Agent.

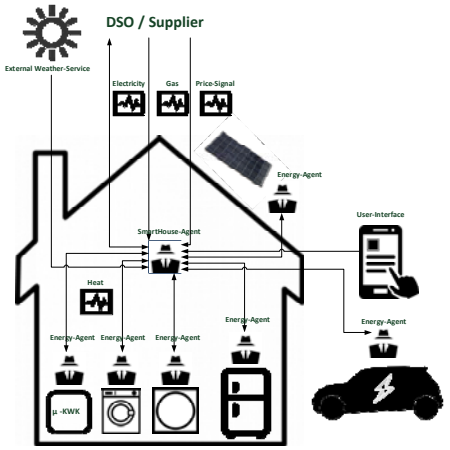


Fig.2: Smart-House Scenario as use case for Energy Agents

As the figure shows, all considered devices are equipped with Energy Agents that include a corresponding EOM for the specific technical system. A concrete example of a single EOM will be shown further below with an Electrical Vehicle.

In detail the dryer and the washing machine were modelled as “smart” batch processes. Such process allows to shift the actual start of the selected program by staying in a waiting operating state. Knowing the energy consumption of the device that is given by measurements of previous runs, the energy usage can be shifted into a designated time range.

The fridge was modelled as repetitive system whose main task is to hold the inner temperature within a range of 4 - 6°C. By varying these temperatures, using them as set-points for the internal controller of the fridge, the Energy Agent is able to increase or decrease these temperatures by ±1°C, in order to flexibilize the energy consumption.

Since the energy production of the photovoltaic plant basically depends on the solar radiation that occurs at the specific day, the power production was just statically considered for the experimental setup. Needed information were taken from historically data for the specific day. In a more elaborated setup the EOM could also be used in order to calculate energy production based on weather information that could be determined by forecasts.

For the µCHP, the manufacturer information of a *Vaillant µCHP ecoPower1.0* was modelled with EOM. Here four operating states were defined that are a standby, an acceleration phase, a normal operation and a shutdown that is followed by the standby of the system again.

In the case of the Demand Side Management scenario the Smart House Agent plays the central role in all optimization issues. It collects and processes all relevant information from every subordinated Energy Agent, like all individual EOMs as well as the actual states of the systems. Thus, the Smart House Agent acts as an aggregator for the whole house. With the aim to optimize energy costs, especially also by first using as much energy as possible from the local energy sources, the optimization approach is carried out for a whole day. In the Demand Response scenario the role of the Smart House Agent was restricted and simplified. Here it basically forwarded external price signals or needed external information, like the outside temperature, needed as forecast from an external weather service. The actual optimization approaches were implemented in each of the Energy Agents.

To demonstrate the efforts in defining a system specific EOM, Figure 3 below shows one exemplary Energy Option Model for the electrical vehicle (EV) used.

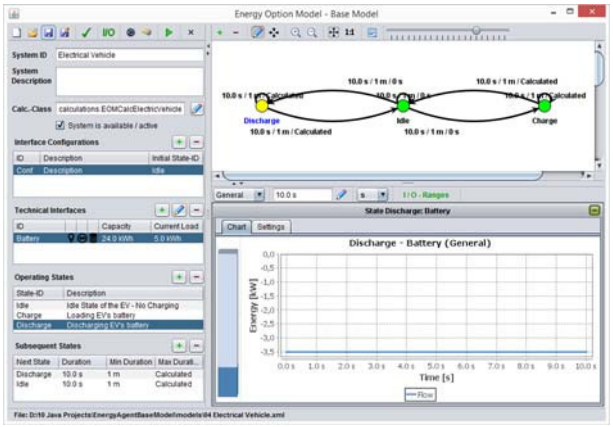


Fig.3: Base Model of an Electrical Vehicle as example

In general, a single technical system is described by the connections to an energy carrier-dependent network that are described by Interface Configurations and Technical Interfaces to the corresponding network (e.g. an EV can differently be connected to an electrical network). If a storage is available, as this is the case for an EV, a system can be defined with one storage per energy carrier. The Image above shows as an example that the used EV has a storage capacity of 24 kWh and a current load of 5 kWh. Beyond that, the image shows the graph of the possible operating states for the EV and possible subsequent states. Here, the operating states Idle, Charge and Discharge are defined. Additionally, for each of this state the actual energy flow are defined, as for example the 3.5 kW electrical output when discharging the EV battery. Figure 4 below shows the settings for the evaluation of a single technical system.

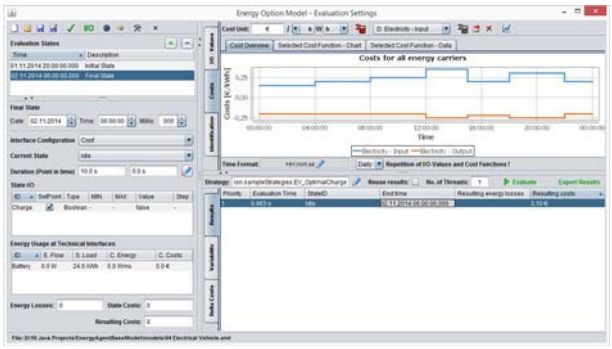


Fig.4: Evaluation settings and result for a single technical system

On the left hand side, the current and the wished technical system states can be defined for the evaluation period. Here the EV was connected to the network at 8 p.m., while the charging process should be finished at the next morning at 6 a.m.; the wished storage load was set to the maximum storage capacity of 24 kWh. On the right side above, the specific cost functions for each energy carrier and for a specific direction (input or output) can defined. It can be seen that the consumption is defined with positive cost values, while the production or feed-in is defined with negative cost values. Developing an individual evaluation strategy a tailored algorithm can be developed and used in order to produce an optimised schedule for the actual technical system. Results can look like the charts shown in Figure 5, where the storage load and the costs over time are presented. In fact, the evaluation strategy used has the aim to cost optimally charge the EV. For this a quite simple algorithm was developed that uses the cost information (shown in Figure 4) and the demand of electrical energy that is given by the goal state within the specified time range. By a simple search of the most inexpensive time ranges for the charging operating state and a Greedy based decision process, the cost optimal schedule for the EV could easily be determined.

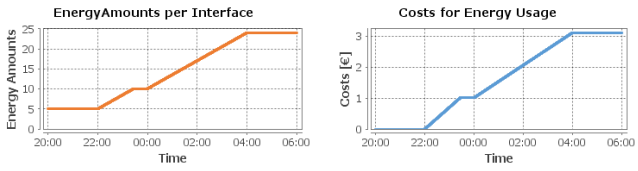


Fig.5: Cost optimized charging of an Electrical Vehicle

As general approach for an evaluation or the optimization of a single technical system, the framework of the EOM provides a graph-based decision system. Based on the time discretization of operational states, the decision system

allows to fully describe technical systems states in regard to (possibly predicted) input and output information, as well as to energy flows at network interfaces. Further, storage loads, transferred energy amounts, energy losses and price signals can be considered for the preparation of a schedule, while the actual optimization problem can individually differ. Figure 6 below shows the so called differences graph of this approach.

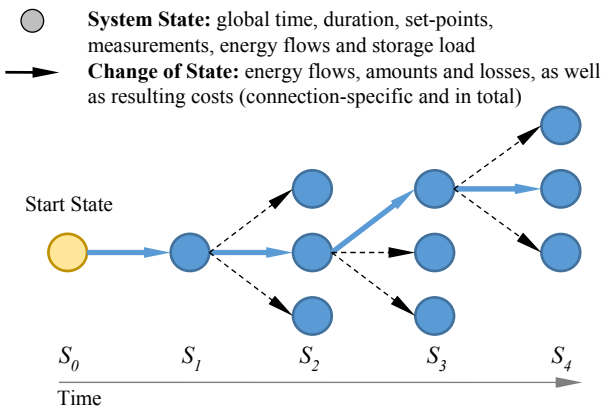


Fig.6: Differences graph for evaluation strategies

The graph will first be created by the current system state S_0 that is defined with the start state for the evaluation process. Since an operating state that is defined by the base model of the EOM describes also the time range for this state, the subsequent state S_1 can be determined with the end of this operating state. With this, energy flows per interface and energy losses that were transferred or produced within this period, as well as the new storage load and thereto corresponding costs will be calculated. After that, and derived from the graph of possible operating states that is defined with the base model of the EOM, the possible subsequent operating states can be determined over time. In order to decide which system states is to be used for a schedule, a decision must be taken at each of the nodes defined by the differences graph. To support informed decisions, the framework of the EOM provides the information mentioned above (energy flows etc.) as pre-calculated set to the decision making process that is located within an individual strategy for an evaluation. Since each technical system might require different background information for such decision processes, the framework of the EOM provides an adaptive programming interface that is named as (Abstract) Evaluation Strategy and that allows to create individual solutions.

The aggregation method for several technical systems that is provided by the EOM and that was used within the Smart House Agent is realized in the sense of a system of systems. Analogously to a single technical system, the aggregator will first be considered as a technical system, summarizing the sub systems in regard to interfaces by energy carrier and by add up storage capacities. In more complex scenarios, were wider areas than a house are to be considered, an aggregation may additionally require a network calculation. Similar to the Evaluation Strategy for single technical systems, a strategy for aggregated technical systems can individually be designed. In contrast to single systems, decisions must be made for each subordinate system, so that several decision graphs will concurrently be used during an evaluation of aggregated technical systems.

What was shown in the illustrations above as user dialog can also be used “head-less” within an Energy Agent. In this case it is the task of an Energy Agent to get and provide the needed base information for a single technical system or an aggregation of technical systems (e.g. cost information) and start an appropriate evaluation process, if required.

For the centrally controlled Smart House example, a first heuristic approach was chosen for the decision making process that discretized again the assessment parameter that in turn corresponded to the optimization goal: the price for energy. More concrete: a high price was rated with lowest priority, while a low price was rated with a high priority. Since we assumed a perfect prediction provided by the EOM’s of the technical systems, no spontaneous re-planning was intended in this scenario. However, we are aware that a re-planning would be necessary if any of the subordinated systems deviates from its prediction. Nevertheless, the state and solution space for the Smart House scenario was already huge and could easily exceed the available computing capacities. The application of the heuristic approach, however, reduced the search and solutions space substantially and with it the time to calculate a result.

Figure 7 below shows the cost functions that were used for the assessment of the energy usage, while Figure 8 presents - as an example - the resulting electrical power flow for the aggregated Smart House scenario.

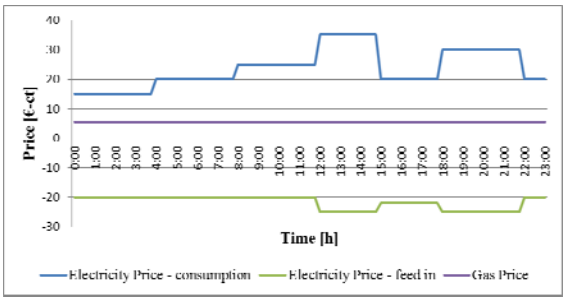


Fig.7: Costs and Revenue Functions by Energy Carrier

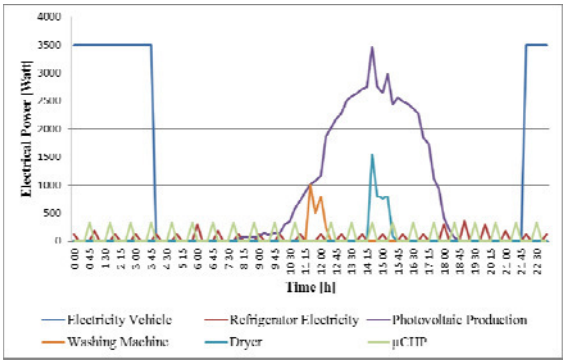


Fig.8: Electrical Power Usage for the aggregated Smart House

Due to the mobility phase, the electrical vehicle was not connected during daytime. Hence, it was charged by night using the cheapest prices from 10 p.m. until approx. 4 a.m. in the morning. The electrical power usage for the washing machine and the tumble dryer was shifted into the power production phase of the photovoltaic plant. Since the µCHP and the fridge are repetitive working systems, these system are providing just small flexibility ranges. It is a task for the future work to improve our approach in order to optimize our first approach.

Overall, it was found that the overall energy costs could be reduced by using the aggregation case and comparing it to the Demand Response is possible. Since in an aggregation the interdependencies of several technical systems can be considered, this result was not surprising us. However, the complexity which we had already to face was.

Much more important for us was, however, to find a way to adaptively describe and connect hybrid technical systems by using the approach of the unifying Energy Option Model. Based on that, more sophisticated optimization approaches will be developed and tested in the future, as well as the number of technical systems considered can hopefully be diversified and increased. Since the developed Evaluation Strategies are based on a graph, several meta-heuristic approaches are conceivable here, as for example ant-based approaches and other.

V. Discussion, Outlook & Conclusion

This paper has presented the concept and the application of Energy Agents that are equipped with a unifying Energy Option Model. Since this model has the capability to describe all relevant types of energy conversion processes, it provides the foundation for the design of new adaptive methods that permit a dynamic aggregation and optimization of groups of technical systems. As a proof of concept an application scenario for a Smart House was presented that compares a demand response case with a case where a Demand Side Management for the house was realized. Based on the Energy Option Model and Agent.GUI, that is already freely available as open source, it is our intention to provide several predefined scenarios as the here presented Smart House scenario in the future.

Moreover, the concept of the Energy Agent will be improved and developed under the project *Agent.HyGrid* in the next years [13]. Here it is planned to close the gap between agent-based simulations and real-world applications in order to produce comparable results for both cases. Thus, a realistic laboratory and test-bed environment will be created for decentralized control solution of Future Energy Grids.

Overall, we believe that urgently standards are required that enable such homogeneous and in particular open developments; for science and for systems used in real applications on-site. We further believe that therefore the concept of a generally accepted Energy Agent with a unified description of the underlying technical system that we call Energy Option Model is the necessary foundation. Together with a refined and commonly accepted concept of an Energy Agent an open Future Energy Grid could be designed that provides the perfect environment for further elaborations and improvements over the next decades. We further believe that such standardization is absolutely indispensable in order to reach the overall objectives, associated with the energy transition or with the planned European Energy Union.

References

[1] C. Oerter and N. Neusel-Lange, “LV-grid automation system - A technology review,” in *PES General Meeting | Conference Exposition, 2014 IEEE*, 2014, pp. 1–5.

[2] E. Bernier, F. Maréchal, and R. Samson, “Multi-objective design optimization of a natural gas-combined cycle with carbon dioxide capture in a life cycle perspective,” *Energy*, vol. 35, no. 2, pp. 1121–1128, 2010.

[3] T. Linnenberg, I. Wior, S. Schreiber, and A. Fay, “A market-based multi-agent-system for decentralized power and grid control,” in *Emerging Technologies Factory Automation (ETFA), 2011 IEEE 16th Conference on*, 2011, pp. 1–8.

[4] C. Derksen and R. Unland, “Hybrid Energy Option Models for Unified Energy Agents,” in *Operations Research, Aachen, Germany*, 2014.

[5] C. Derksen and R. Unland, “An advanced agent-based simulation toolbox for the comprehensive simulation of future energy networks,” in *Smart Grid Technology, Economics and Policies (SG-TEP), 2012 International Conference on*, 2012, pp. 1–4.

[6] D. Nestle and J. Ringelstein, “Integration of DER into Distribution Grid Operation and Decentralized Energy Management,” *Smart Grids Europe*, vol. 19, 2009.

[7] [Online] Available at: http://www.e-energy.de/de/modellregionen.php.

[8] A. Faruqui and S. Sergici, “Household response to dynamic pricing of electricity: a survey of 15 experiments,” *Journal of Regulatory Economics*, vol. 38, no. 2, pp. 193–225, 2010.

[9] I. Pinyol and J. Sabater-Mir, “Computational trust and reputation models for open multi-agent systems: a review,” *Artificial Intelligence Review*, vol. 40, no. 1, pp. 1–25, 2013.

[10] O. Boissier, J. F. Hübner, and J. Simao Sichman, “Organization Oriented Programming: From Closed to Open Organizations,” in *Engineering Societies in the Agents World VII*, vol. 4457, G. M. P. O’Hare, A. Ricci, M. O’Grady, and O. Dikenelli, Eds. Springer Berlin Heidelberg, 2007, pp. 86–105.

[11] M. Wester-Ebbinghaus, “Von Multiagentensystemen zu Multiorganisationssystemen - Modellierung auf Basis von Petrinetzen,” Universität Hamburg, Von-Melle-Park 3, 20146 Hamburg, 2010.

[12] C. Derksen, T. Linnenberg, R. Unland, and A. Fay, “Unified Energy Agents as a Base for the Systematic Development of Future Energy Grids.,” in *MATES*, 2013, vol. 8076, pp. 236–249.

[13] C. Derksen, T. Linnenberg, N. Neusel-Lange, and M. Stiegler, “Agent.HyGrid: A seamless Development Process for agent-based Control Solutions in hybrid Energy Infrastructures,” in *SmartER Europe - E-world energy & water, Essen, Germany*, 2015.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 265–271
DOI: 10.15439/2015F227
ACSIS, Vol. 6

A Day-ahead Centralized Unit Commitment Algorithm for A Multi-agent Smart Grid

Salam Hajjar
Universidade Federal do Rio de Janeiro, COPPE, Ilha do fundo, Rio de Janeiro, 21.945-970 RJ, Brazil

Antoneta Iuliana Bratcu*
University of Grenoble Alpes
GIPSA-lab
Saint-Martin d'Heres, France

Ahmad Hably
University of Grenoble Alpes
GIPSA-lab
Saint-Martin d'Heres, France

***Abstract*— Renewable energy resources like wind and solar have become an effective factor in the energy production on the planet as they are inexhaustible renewable resources. However, they are very intermittent and their output cannot be predicted certainly. In this paper an algorithm of unit commitment within a power grid integrating wind and photovoltaic production units is proposed in a centralized approach that takes into account provisional data about the renewable energy production. Here the unit commitment problem is stated as a power demand coverage problem with some prespecified merit order list. A multi-agent architecture is proposed to facilitate the message exchange and easy addition and deletion of agents in the grid. This architecture is flexible and easy reconfigurable as it can provide solutions under assumptions of a decentralized approach. An implementation using JADE platform is presented in this work. The system is tested using real-data sets from an existent energy transport network in France (RTE). The results based on different operating conditions show the economic sense of the proposed strategy.**

I. Introduction

The unit commitment (UC) problem aims at finding the least-cost dispatch of available generation resources to cover the load demand in an electrical grid. Production units' availability and setpoints are key decision issues in UC problem. Short Run Marginal Cost (SRMC), which is defined as the pure cost of electrical power production, is supposed to be the market-based price of electricity. However, costs of power generation evidently exceed this price, thus, power generators offer generation prices of their choice [1]. Obviously the power load depends strongly at each spacial zone on the period of the day and the season of the year. For example, in France a reasonably reliable demand forecast, as well as the list of available power units with their updated technical constraints and economic characteristics, are available each day by 4 p.m. Appropriate production schedules must then be computed, to be sent to the local units by 5 p.m. [2]. In the last years, renewable energy resources such as photovoltaic and wind turbine plants have been considered as a nonnegligible part of the energy offer since these resources are inexhaustible and their production cost is low compared to the conventional energy resources. However, uncertainty surrounds the availability of these renewable energy resources. The grid behavior becomes uncertain and the unit commitment problem must ensure the power demand covering without the need of a precise knowledge about the renewable energy units output.

A solid mathematical foundation has been built to help better understand the stochastic energy constraint and the inherent correlation between quality of energy and the uncertain energy supply [3]. Recently the domain of multi-agent systems has achieved a real progress, it has been used to solve spatially-distributed and open problems. Power grids have a spatially distributed architecture and can be considered as systems made up of agents, some of which are energy producers, some other consumers acting on the energy market. For this, multi-agent architectures may be built to tackle different problems in power grids, such as the unit commitment problem.

In this paper, a centralized decision making algorithm is proposed to solve the UC problem at the level of an arbitrarily defined spacial geographical entity, region, or country, by using the advantage of a multi-agent architecture. The unit commitment problem is here focused on power demand coverage; to this end, it is based on a prespecified merit order list that can result from taking into account particularities of the considered geographical entity. The multi-agent architecture provides data exchange and avoids the problem of incompatibility between different types of software communicating agents. Assuming that agents are collaborative, the centralized decision provides a certain level of uniformity and coherence in controlling the grid.

This paper is organized as follows. Section II synthesizes some of related works. Section III shows the multi-agent architecture considered for the UC algorithm implementation. Section IV illustrates the proposed unit commitment algorithm. Section V discusses the results of applying the algorithm on real data scenarios. Finally, Section VII concludes the presented work and provides some perspectives.

II. Review of Related Work

Decision making within power systems has been largely treated in the literature with different points of view. In [4], a distributed price-based control method is proposed. The controller takes its decision information from its neighbors. A day-ahead UC algorithm has been proposed in [5], where a decision making is achieved by a two-level decentralized

salam@poli.ufrj.br
*Corresponding author
antoneta.bratcu@gipsa-lab.fr
ahmad.hably@gipsa-lab.fr

framework. A low level treats each production unit individually to optimize its benefits. At the high level, coordinators communicate with their neighbors to update the prices by using a distributed method. In [6] energy storage systems (ESS) are studied as an alternative to provide flexibility in power system operation. The integration of a grid-scale ESS in short-term operational planning in a centralized cost-based electricity market has suggested easier integration of renewable power resources as dispatchable units into a power generation grid. The rolling horizon strategy has been proposed in [7] as base of an energy management system for a renewable-energy-based micro-grid. In such strategy the control decision results iteratively from solving at each step a mixed-integer optimization problem. In order to provide on-line setpoints for each generation unit and signals for consumers based on a demand-side management mechanisms, forecasting models are used. This work treats globally the different agents for the benefit of the entire grid instead of concentrating on the individual benefit of each single producer.

An intelligent energy network system based on multi-agent systems has been proposed in [8]. It illustrates the advantage of using multi-agent structure over the centralized approach. Three types of agents – producers, loads and market – contribute at solving the UC problem and providing the best setpoints planning according to information exchanged between the agents regarding the day-ahead consumption and production prevision. The work has been implemented on JADE multi-agent framework. This paper relies on the same idea, proposed in [8], of reducing the amount of information exchanged between agents. The two approaches differ at the level of computing the setpoints and treating the spot market.

III. Multi-agent architecture

Data exchange between agents may be simplified; thus, agents communicate between each other only the necessary information for the decision making. A multi-agent system is extendable open and open-ended system, i.e, agents can leave and new agents can join the system randomly, so that a theoretically infinite number of structures can be obtained. Dynamical behavior of agents makes that network management to be quite delicate process. Thus, some assumptions on the system can be adopted in order to facilitate the network's management. In this work two kinds of assumptions are made, which are categorized and listed as follows:

- Assumptions on the agents' behavior:
  - Each agent must provide its status. If it is active, then its status is ON, otherwise it is OFF.
  - An agent must maintain its status and the information exchanged with the other agents until the next cycle of data exchange.
  - Agents are supposed to provide reliable information on their production, consumption, prices and costs.
- Assumptions on the grid architecture:
  - The grid allows the integration of agents where the category is already recognizable by the grid. Agents of unknown or with undefined types will not be taken into account in the decision taking algorithm.
  - The grid allows and does not limit the communication between its local agents. Agents can communicate with other agents out of the grid if assumptions on their behavior are respected.

The considered multi-agent grid is built up of five categories of agents representing the main actor types in a real power grid: (1) producer, (2) consumer, (3) storage unit, (4) external market and (5) controller or dispatcher. Since the focus is here on the decision-making process performed by the controller, the internal behavior of producers, consumers, storage units and external market is not detailed in the multi-agent grid model considered in this work. Thus, agents are described mainly by their role on the market, as explained in the following set of assumptions about the energy market:

- Producer: This agent is an energy producer that offers its production on the market. Five types of producers are considered here: (1) Nuclear plant, (2) Hydraulic turbine, (3) Gas turbine, (4) Wind turbine plant and (5) Photovoltaic plant (PV). Any producer agent has a five-element profile: **producer_type**,**energy_produced**, **price**, **starting_cost**,**state**.
- Consumer: This agent represents any energy consumer which can be an industrial one, a residential one, etc. A consumer agent has a profile composed of only one element, **consumption** which represents energy to be consumed;
- Storage unit: This agent is an aggregated energy storage unit for the entire grid. This unit can store energy and/or provide energy to grid depending on the storage capacity and availability. This agent has a three-element profile: a limited **storage_capacity**, **store_energy** and a **storage_cost**. The store is considered unavailable if it is empty, i.e, the **store_energy** equals 0 MWh;
- External market: This agent represents an image of the market outside the considered grid. The grid can buy energy from the external market if the local producers cannot cover the consumers' needs. In this case, market can be seen as a source of energy. Extra energy produced locally can be sold to the external market instead of storing it, in case of financial benefit to the grid. It is supposed that external market can always absorb the extra energy. The decision to store the energy or sell it is taken by the UC algorithm. The market's profile consists of four elements: **offered_energy**, **selling_price**, **purchasing_price**, **purchased_energy**.
- Controller: This agent is responsible of providing a day-ahead setpoint, planning of agents needed to cover the demand, prevision of potential extra energy production, decision whether to store or to sell the extra energy and the benefit of selling to external market. The controller profile has no elements since its aim is only to provide decisions upon running the UC algorithm based on data sent by the agents.

Each agent has specific tasks to perform and a profile which determines its characteristics. The agent by itself can be seen as representing many instances of its type. Agents communicate with each other through messages which contain only the necessary information for taking the UC decision. These messages in their simplest form contain only the values of the agent profiles' elements. The decision is taken in a centralized manner by the controller. The controller is assumed to be a trusted agent; thus, other agents can communicate to the controller true values of consumption, production and prices. The relation between all agents is represented in Figure 1.

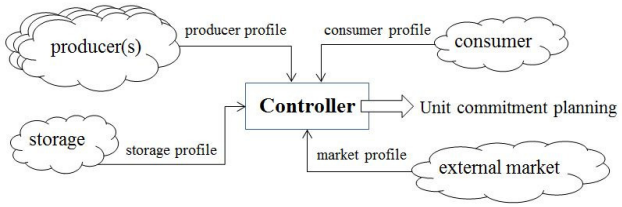


Fig. 1. Smart grid communicating agents.

IV. Unit commitment algorithm

In this section a day-ahead unit commitment algorithm with a decision horizon of 48 periods (30 minutes each) in the day-ahead market is considered. The algorithm has two main goals: (1) to cover the consumption by the local production using the most trustful agents available; in this way consumer needs can be met through the offers proposed by the most available producers who make the most interesting offers; (2) to maximize the production benefit by selling the extra produced energy, if any, to the external market.

A. Assumptions

In order to satisfy the consumption needs, the controller has direct connection with producers, storage and the external market. Producers, consumers and storage send messages to the controller regarding their production, consumption needs and offers information, respectively. The controller decides, depending on the received information, which producers must be called to cover the consumer needs and in which order. Dispatching order will depend on producers dispatchability and size. From this point of view, renewable energy units have low dispatchability because of their intermittence and unpredictability. Therefore, they will not be sollicited as main producers, but as so-called backup units. Among the different types of producers mentioned in Section III studies show that the nuclear plants come in first place in energy production as their production is fully dispatchable and may reach thousands of MW depending on the plant size. For example, the production of the nuclear central in Rhône-Alpes region in France can range from 4000 MWh to 12000 MWh. Hydraulic turbines come in the second place as they can produce energy that ranges from 1000 MWh to 4000 MWh. Gas turbines come in the third place as their production reaches only to few hundreds of MW. Wind and photovoltaic plants have the lowest sollicitation level as their production capacity is strongly related to the season of the year, the hour of the day and the geographic location
In order to propose a day-ahead planning for the agents in the system, the controller, at each sampling period, calls the available agents whose status is ON. The algorithm calls first the fully dispatchable production units, then the backup ones. At first place it calls the nuclear producer with the entire amount of produced energy. It compares the consumption demand with the energy produced by the nuclear agent. If the nuclear agent cannot cover alone the demand, then the algorithm calls the hydraulic turbine agent. It calculates the sum of the energy produced and compares it to the consumption demand. If the demand remains unsatisfied at this point, the controller calls the storage. Since the storage does not have a starting cost it becomes practically cheaper than the gas turbine producer. The controller continues calling agents and check power balance. It calls the gas turbine agent, then the PV agent and finally the wind turbine agent. It is considered judicious for the PV plants to come in priority before the wind turbine agent since PV plants are cheaper to start, although they are not available the whole day. Such choice is confirmed in [9]. At each step, if the accumulated value of produced energy exceeds the consumption demand, the controller stops calling agents and calculates the amount of extra energy as follows:

$$\textbf{extra\_energy = accumulated\_energy - consumption} \quad (1)$$

In case extra energy is produced the controller decides either to store this energy or to sell it to the external market. The decision to sell or store relies on the following reasoning:

If $(extra\_energy \times market\_selling\_price(p)) > [extra\_energy \times market\_selling\_price(p+1) - extra\_energy \times stock\_cost(p)]$

then sell the $extra\_energy$ at period p

else store the amount of energy regarding the storage capacity to be sold at period $p+1$ and sell the remaining at period p

Since this decision is strongly dependent on the external market price, it is quite difficult to predict it, as the external market price oscillates depending on the spot market price.

B. Decision-making strategy

The 10-step algorithm flowchart is given in Figure 2. At step 1, the controller collects the required data and reads the messages sent by the available agents in the network. At step 2, the controller initializes two lists: (1) a list "$available\_agents$", which contains the activated agents with their profile values, (2) the list "$called\_agents$", initially empty, which will contain, by the end of the algorithm, the group of agents which satisfy together the consumption demand. In case of an absent agent, its profile will be replaced by $null$ in the list and values of its elements will be replaced by zeros. At step 3, the agents in the

list "$available\_agents$" are sorted, by type. Each agent is given a priority degree, so that the controller calls the agents in their order. At step 4, the controller verifies if the consumption demand is satisfied. Obviously, at the first cycle of the algorithm the demand would not be satisfied, since no agents are yet called. At step 5, the controller calls an agent from the list "$available\_agents$" according to the priority order. At step 6, the controller updates the list of called agents. At step 7, the satisfied demand is updated depending on the energy offered by the last called agent. This update is computed as follows:

$$satisfied\_demand := satisfied\_demand + agent\_energy \quad (2)$$

At step 9 the extra energy production is computed as follows:

$$extra\_energy := consumption - satisfied\_demand \quad (3)$$

If extra energy is produced then the controller goes to step 9, otherwise, the algorithm returns to step 4, continues towards the end and provides the proposed UC planning. At step 9, the decision whether to sell the extra energy at the current period p or to store it, is taken knowing that the storage agent has a limited capacity, it is not sure that the entire extra energy can be stored. If it cannot be stored, extra energy is sold to the external market at period p whatever its financial income is. The algorithm arrives to its end at step 10, which is reachable either from step 4 or from step 9. If the end is reached from step 4, this means that the exact consumption amount can be produced by the called agents and no financial gain is obtained. When the end is reached from step 9, an extra energy has been produced and some financial gain is earned. At step 10, the UC plan is proposed for the period p. It consists of a list of called agents with a detailed report illustrating the consumption demand, the production that covered the demand, the amount to be paid to the called agents, which is the sum of their individual prices, the decision taken about the selling to the external market or storing and the selling income. It is supposed that the controller requires from the storage unit only the needed amount of energy to cover the demand. Thus, at any period p, if the storage contains the energy needed to cover the consumption, then no extra energy is necessary.

V. Case studies

The proposed UC algorithm is applied on two case studies. The first one represents the consumption and production in a normal year day in France, the 15^{th} of May 2014. The second is 25^{th} of December 2013, a winter day during holidays where the consumption is supposed to be different from typicall. The day-ahead consumption and production prevision data is picked up for the Rhône-Alpes and Provence-Alpes-Côte-d'Azur from the website of French electricity transport grid (Réseau du transport d'électricité - RTE, http://www.rte-france.com/en/eco2mix/eco2mix-donnees-regionales-en).

A. JADE implementation model

In this work JADE (Java Agent Development Framework) platform [10] is employed to implement the UC algorithm.

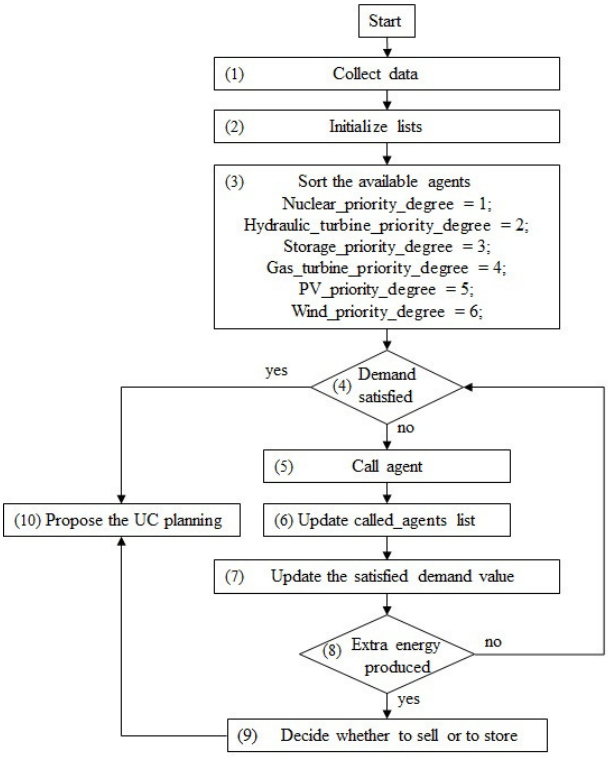


Fig. 2. Overview of decision-making flowchart.

JADE is a programming framework that significantly facilitates the implementation of agent-based applications in compliance with the FIPA [1] specifications. Five producer agents (nuclear, hydraulic turbine, gas turbine, wind and PV), a grid-level storage unit, a consumer which represents the aggregated consumption of the entire grid in the considered region and the external market are modeled by software agents. The behavior of each agent is described by a set of arrays containing the values of the agent profile's elements. The controller agent is a decision-making software agent who runs the UC algorithm explained in section IV. The agents – producers, consumer, external market and storage – send messages to the controller containing the integer values of their individual profiles, i.e, the previsions of production, consumption and prices for the next day. The controller computes the setpoints for 48 time intervals and proposes the agents' commitment planning for the next day.

B. Discussion of results

For May 15^{th} 2014 scenario Figure 3 illustrates the energy production predicted for PV plants, wind turbine plants and gas turbines one day ahead. Wind power prevision is based on the slowly variable, seasonal component of wind speed, so a-half-an-hour sampling time is sufficient for good prediction accuracy. Note that the production value of these producers ranges from 0 MWh to 220 MWh at its highest value produced by the PV agent at its peak of production.

[1]Foundation of Intelligent Physical Agents

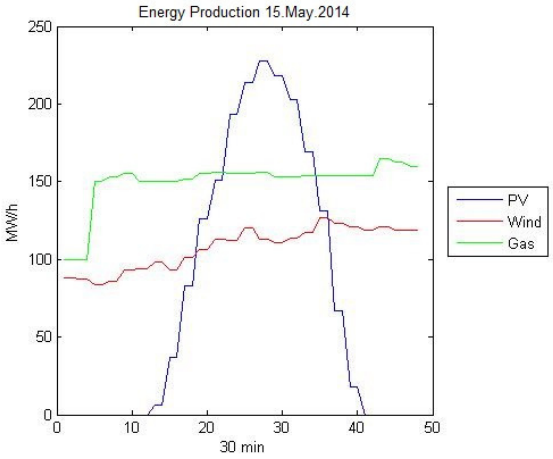


Fig. 3. PV, wind, and gas production prevision for May 15th 2014.

Figure 4 illustrates prevision of consumption needs and production of the different types of producers. PV, wind and gas turbine productions have been represented as an aggregated curve called backup energy. Logarithmic scale in y axis is preferred since the backup energy is much smaller compared to those of other agents. The consumption at that day ranged from 10500 MWh to 14600 MWh. Note that production of nuclear and hydraulic turbine agents is much higher than production of the gas turbine, PV and wind agents. The UC algorithm has been applied to this

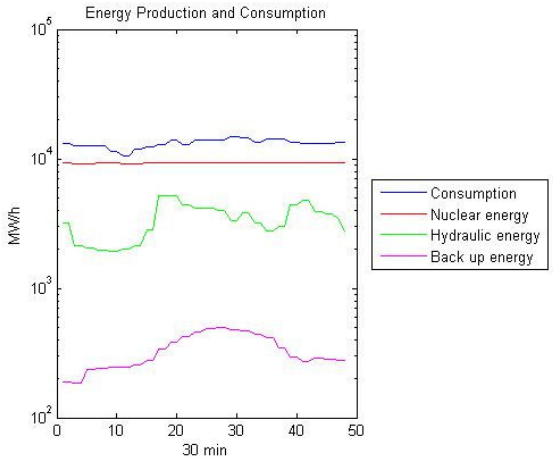


Fig. 4. Consumption and production prevision for May 15th 2014, logarithmic scale.

consumption and production scenario. Figure 5 illustrates the consumption cover realized by the consumers, and the extra energy produced at periods p_i where [2] $i = \{[8-11], [14-25], [38-45]\}$. Here also, the backup energy cummulates PV, wind and gas turbine productions and y axis values are in logarithmic scale. A zoomed view over the extra energy produced is shown in Figure 6. The algorithm decides according to (2) and the market price evolution in Figure 7 whether to sell the extra energy at period i or to store it. The algorithm indicates for the scenario of 15^{th} of May that the immediate selling of extra energy is beneficial compared to storing the energy. Computation of the selling income shows that it reaches its highest value at periods 18 and 19, which correspond to the peak of market price and the peak of extra energy production. However, this prevision of income cannot be taken as systematically predictable behavior since neither the market price, nor the extra energy production are constant or systematically repetitive.

[2]from period 8 to 11, from period 14 to 25 and from period 38 to 45

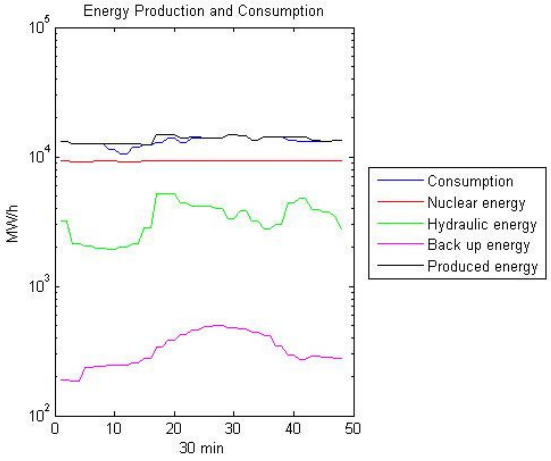


Fig. 5. Consumption needs covering on May 15th 2014, logarithmic scale.

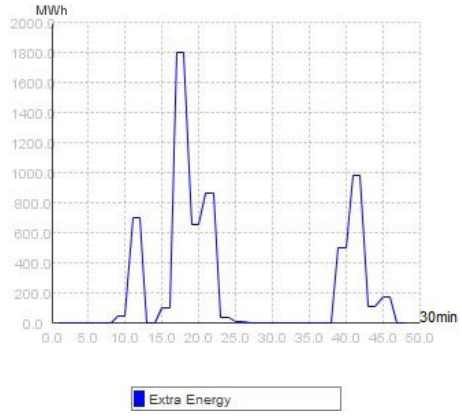


Fig. 6. Extra energy produced on May 15th 2014.

C. Results comparison

We focus in this discussion on the effect of the PV and wind energy resources on the grid. We argue the financial advantage of integration of these resources in a large-scale grid such as the one considered in this work. PV and wind production prevision data for the 25^{th} of December 2013 are shown in Figure 8. Note that the PV production peak dramatically decreased from 210 MWh in May to 32 MWh in December. Similarly the wind production peak decreased

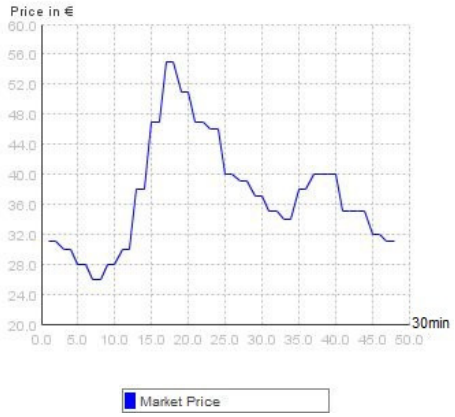


Fig. 7. Spot market price on May 15th 2014.

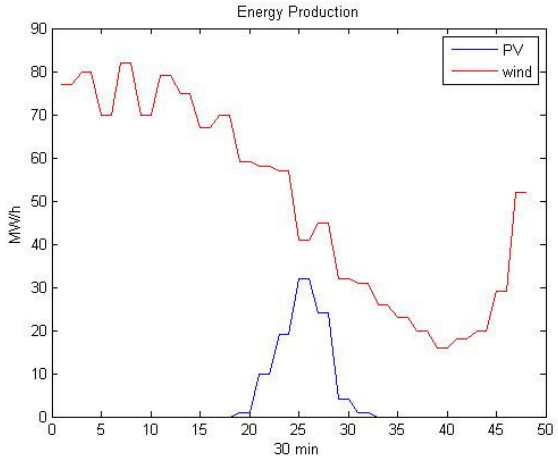


Fig. 8. Wind and PV produced energy on December 25th 2013.

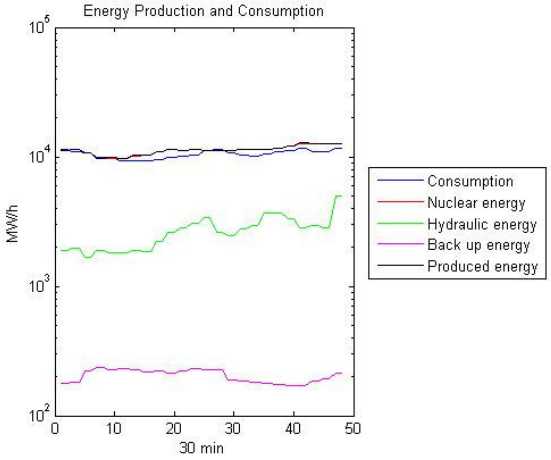


Fig. 9. Consumption covering on December 25th 2013, logarithmic scale.

from 163 MWh in May to 82 MWh in December. The algorithm results on the scenario of 25^{th} of December, illustrated in Figure 9, show on one hand that the considered UC algorithm achieves covering the consumers' needs by calling the external market at most of the day time. The backup energy cummulates PV, wind and gas turbine productions and y axis values are in logarithmic scale. On the other hand, a very small amount of extra energy is produced at periods 10 and 25 which are the peak periods of wind and PV, respectively. This confirms that the renewable energy plants can be integrated into the power grid as backup agents. When their energy is not necessary to cover the power demand, it can be sold to the external market. Practically this might be much effective financially in Spring than in Winter.
Indeed, the considered context – characterized by use with predilection of nuclear energy – render the renewable energy sources dependent on the external market price. This situation does not foster development of renewable small-power sources as grid units, unless they are made fully dispatchable. To this end, a possibility is their association with storage units and implementation of an effective local power flow management.

VI. Scalability and Performance

The scalability of the proposed algorithm depends on several factors: the messages exchange time, the size of data within messages and the number of communicating agents. To evaluate the scalability of the platform, one can start with a small number of communicating agents, for example, the controller, the customer and two power providers, then enlarge the network by increasing the number of power providers. One can note that the computation time grows when the number of network agents increases. This results from the increase in the number of messages exchanged between agents and the time needed to extract the data out of the messages. The performance of the algorithm depends strongly on the implementation environment, i.e, the performance of each agent's machine and the network speed. In this work, the used implementation environment is one Dell personal computer, where Jade 3.2 and Eclipse IDE for Java developers are installed. The algorithm has been run off-line since JADE platform played the role of multi-agent network. Thus, the computer characteristics strongly influence the algorithm performance. We indicate here that the run time of the studied example was about 0.5 seconds. Table I provides the characteristics of used PC.

VII. Conclusion

The work presented in this paper proposes a centralized unit commitment algorithm, implemented on a multi-agent architecture. The algorithm achieves to cover the demand of consumers in a power grid by taking into account daily-variation-based previsions of renewable energy production, mainly PV and wind turbines. It illustrates the advantages of using a multi-agent architecture in simplifying the exchange of messages between agents. The proposed algorithm is applied to real data retrieved in Rhône-Alpes and Provence-Alpes-Côte d'Azur in France. Analyzing how the proposed algorithm optimizes the schedule of power plants with respect to total cost is a further issue. Such analysis may be

TABLE I
IMPLEMENTATION MACHINE'S CHARACTERISTICS

Model	Dell Inspiron
Processor	Intel core i5 CPU M560@2.67 GHz x 4
Total memory	3.7 GB
Operating System	64-bit Debian
OS version	7.3 (Wheezy)
JADE version	3.2
Eclipse version	IDE for Java developers
JAVA	Sun SDK 1.4

useful to assess effectiveness against other methods for solving the unit commitment problem in its classical formulation. The proposed multi-agent architecture is flexible to accommodate different centralized decision making strategies. This architecture can easily be modified to accommodate more complex agent scenarios and agent interaction mechanisms, energy market mechanisms and decision-making strategies, in order to put into light different patterns of behavior within a decentralized context, that are difficult to predict.

ACKNOWLEDGMENT

This work was achieved in the framework of the Smart Energy project funded by Grenoble Institute of Technology in Grenoble, France, to whom authors are kindly grateful.

REFERENCES

[1] M. Lively, "Short run marginal cost pricing for fast responses on the smart grid," in *Procs. of 2010 Innovative Smart Grid Technologies,* January 2010, pp. 1–6, DOI 10.1109/ISGT.2010.5 434 779.

[2] L. Dubost, R. Gonzalez, and C. Lemar´ echal, "A primal-proximal heuristic applied to the french unit-commitment problem," *Math. Program.,* vol. 104(1), pp. 129–151, 2005, DOI 10.1007/s10107-005-0593-4.

[3] W. Kui, J. Yuming, and D. Marinakis, "A stochastic calculus for network systems with renewable energy sources," in *Procs. of 2012 IEEE Conference on Computer Communication Workshops,* March 2012, pp. 109–114, DOI 10.1109/INFCOMW.2012.6 193 470.

[4] A. Joki´ c, M. Lazar, and P. van den Bosch, "Price-based control of electrical power systems," in *Intelligent Systems, Control and Automation: Science and Engineering* (vol. 92 Intelligent Infrastructures), R. R. Negenborn and Z. Lukszo and H. Hellendoorn (Eds.), 2010, pp. 109–131, DOI 10.1007/978–90–481–3598–1 5.

[5] L. Mingyang and P. Luh, "A decentralized framework of unit commitment for future power markets," in *Procs. of 2013 IEEE Power and Energy General Meeting,* July 2013, pp. 1–5, DOI 10.1109/PESMG.2013.6 672 790.

[6] C. Suazo-Martinez, E. Pereira-Bonvallet, R. Palma-Behnke, and X.-P. Zhang, "Impacts of energy storage on short term operation planning under centralized spot markets," *IEEE Trans. on Smart Grid,* vol. 5(2), pp. 1110–1118, 2014, DOI 10.1109/TSG.2013. 2281828.

[7] R. Palma-Behnke, C. Benavides, F. Lanas, B. Severino, L. Reyes, J. Llanos, and D. Saez, "A microgrid energy management system based on the rolling horizon strategy," *IEEE Trans. on Smart Grid,* vol. 4(2), pp. 996–1006, 2013, DOI 10.1109/TSG.2012.2231440.

[8] S. Abras, C. Ki´ eny, S. Ploix, and F. Wurtz, "Mas architecture for energy management: Developing smart networks with jade platform," in *Procs. of 2013 IEEE Intl. Conference on Smart Instrumentation,* Measurement and Applications, November 2013, pp. 1–6, DOI 10.1109/ICSIMA.2013.6 717 913.

[9] M. Zipf and D. Most, "Impacts of volatile and uncertain renewable energy sources on the german electricity system," in *Procs. Of 2013 10th Intl. Conference on the European Energy Market,* May 2013, pp. 1–8, DOI 10.1109/EEM.2013.6 607 397.

[10] F. Bellifemine, G. Caire, and D. Greenwood, *Developing Multiagent Systems with JADE.* New York, USA: Wiley, 2007. DOI 10.1002/ 9780470058411.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 273–280
DOI: 10.15439/2015F330
ACSIS, Vol. 6

Evaluation of distributed multi-agent Energy Management System – cost calculation

Weronika Radziszewska
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
Email:Weronika.
Radziszewska@ibspan.waw.pl

Jörg Verstraete
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
Email: Jorg.
Verstraete@ibspan.waw.pl

Jacek Wasilewski
Institute of Electrical
Power Engineering,
Warsaw University of Technology
Warsaw, Poland
Email: Jacek.
Wasilewski@ee.pw.edu.pl

***Abstract*—An Energy management system (EMS) is a concept that spans various possible solutions, ranging from basic implementations over solutions that use simple intelligent computer methods to systems that employ advanced intelligent methods. We designed and implemented an intelligent multi-agent based approach that uses a market mechanism to manage power in the microgrid of a simulated Research and Education Center. In this article, the performance of our system regarding cost optimality is compared against two other solutions: the first is a simple solution that uses predefined profiles that define the use of the controllable sources - this provides an upper limit to the cost, the second is the perfect artificial optimum which provides a lower limit to the cost. The perfect optimum resembles a centrally controlled EMS, with the difference that it does not suffer from the delays of detecting power imbalances. The tests show that distributed approach closely resembles the optimal one.**

I. Introduction

The concept of the microgrid introduces new possibilities regarding more optimal and cheaper use of the renewable power sources and management, not only of supply of power but also of consumption. A microgrid is a small sized grid, equipped with its own power sources, that has a single connection to national power grid. In addition, it can be (temporarily) disconnected from the national power grid. The microgrid allows for more internal balancing of power, which increases safety of power supply, better use of power from renewable sources and allows for reducing peaks in the national power grids. There are many features that discern microgrids from big power systems. The issue has been discussed in detail in [1]. Essential features for functioning microgrids as semi-autonomous power systems include the use of power electronic converters and the use of specific control systems. It also needs the ability to communicate within the microgrid.

Energy management systems (EMS) are systems to control the energy in some area, the energy can be electrical, heat or even cooling systems – only electrical energy is considered in this article. An EMS can be a simple implementation, but it can also employ more advanced computer methods, even involving advanced intelligent methods. The EMSs are the key to smart grids and to microgrids' efficiency. There are many EMS solutions for smart grids, from centralized solutions [2], via hierarchical ones [3] to distributed ones [4]. In this article an evaluation of a distributed multi-agent module of EMS – Short-time Balancing System is discussed.

The Short-time Balancing System was created to manage a very specific microgrid: our test case of a Research and Education Center (REC). The detailed description of REC is presented in next section. The requirements of EMS were to be a fully distributed, agent-based system, whose balancing mechanism includes the a auction mechanism. The general description of the system will be presented in section III.

The algorithms and methods used to optimize the speed of power balancing by the Short-time Balancing System were presented in [5] and [6]. To evaluate the algorithm of balancing it should be compared to a benchmark solution. The fact that the system and REC are simulated, allows for different experiments and configurations. The performance of our system with respect to cost optimality is compared against two other solutions: the first is a simple solution that uses predefined profiles that define the use of the controllable sources - this provides an upper limit to the cost (this solution is described in section VI), the second is the perfect artificial optimum which provides a lower limit to the cost (described in section V). The perfect optimum resembles a centrally controlled EMS, with the difference that it does not suffer from the delays of detecting power imbalances; a true centralized system will have a cost slightly above the optimum, as any system needs time to detect imbalances when they occur and also needs time to adjust to them.

The comparison value is the cost in PLN of a day of operation of the microgrid. The costs are presented in section VII. The last section concludes the article.

II. Research Center microgrid

The aim of the REC is to research new renewable energy and smart grid technologies, as well as spread the knowledge about more environmental behaviors, by organizing conferences, courses and seminars. It is assumed that the center will host a research institute, a conference center and a hotel. The microgrid is a low voltage network (LV, 0.4 kV), connected to a medium voltage supply line (MV, 15 kV) via an MV/LV (15/0.4 kV) transformer substation [7]. The REC is equipped with the following renewable power sources: 2 sets

of photovoltaic panels and 3 wind turbines; it also contains two controllable micro power sources: a gas microturbine and a gas-fueled reciprocating engine. Additionally, the microgrid is equipped with two types of power storage units: a set of accumulator batteries and a set of flywheels.

The microgrid consists of 128 nodes, where a node is a point of the network to which the devices with a similar purpose are connected, it is the smallest unit that is measured and controlled. The number of devices that are connected to a single node varies: some nodes can gather a large number of small devices (e.g. lights in the conference room), others have a single device connected (specially used to provide better insight in specific devices, e.g. gas microturbine). A more detailed description of the microgrid can be found in [5].

The microgrid is equipped with controllable devices. For these, the EMS can request to adjust the produced or consumed power accordingly, within the technical limitation of the device. Such devices are non-renewable power sources (reciprocating engines, gas microturbines), as well as batteries. Wind turbines and photovoltaic panels are considered uncontrollable devices, as the amount of their production depends on the weather conditions and cannot be changed by the EMS. Bigger wind turbines are controllable, but the smaller models used in the microgrid lack this functionality.

III. EMS AND BALANCING ALGORITHM

The EMS developed for this microgrid was made with the idea of a complex approach to energy management. The full EMS includes various modules: from demand side management by scheduling events that consume significant amounts of power to calculating the reliability indices for continuity of power supply. The details of the system are presented in [8]. One of the components of the EMS is a Short-time Power Balancing system, which is responsible for changing the operating point of controllable devices in the microgrid. Such control is directly influencing the cost of operation of the REC. The Short-time Balancing system is a distributed and multi-agent system, it is not dependent on a central control point and allows for distributed calculation, which increases the reliability of the solution. The multi-agent paradigm was considered to be very useful in this situation: agents can be given the necessary intelligence to make a decision, and communication between agents is one of the main aspects in a multi-agent system. It also introduces a parallel operation, which on one hand increases the speed of balancing, but on the other hand can cause the unwanted behavior and race conditions. A more detailed description of this system is presented in [6]. The Short-time Balancing system's aim is to optimize the operation of the microgrid.

The choice of which controllable source to use is made by means of relative cost, which not necessarily has to reflect the real cost. If it reflects the real cost the cheapest sources will be preferred, other cost assignments can result in a power balancing scheme that behaves differently. The most intuitive purpose is to minimize the cost of operation, but other purposes can be considered. Maximizing the use of clean power sources could be a second criterion. A third purpose can be shaving the peaks of power usage: in power networks, peaks of power usage pose a problem, as the grid has to be able to cope with those peaks. Microgrids, with their own sources and storage, can help in shaving the peaks. The criteria for optimization considered here is a combination of cost minimization, environmental optimization and optimal usage of the resources present in the microgrid.

Considering the aim of REC and the reasons it was designed, it is assumed that the exchange of power between the external power grid and the microgrid should be minimized. It is not always the case that the power from external (national) power grid is more expensive or more polluting (the ecological aspects are also considered), but considering that the prices of power in external power grid tend to grow and that the microgrid has an ability to work in island mode (disconnected from external power grid and balancing the demand and supply at every moment), the scenario of limiting the exchange of power seems very desired. In this case we assume that selling power to the external power grid brings minimal or no income and buying power from the external power grid is more expensive than producing it internally. The Short-time Balancing System is made in a way that the relative prices (and the same order of the preferred producers) can be easily changed and the system will adjust itself to different model of balancing.

The Short-time Balancing System does not manage the operating point of power consuming devices. The demand side management in the system is realized by the Planner, a module that schedules tasks that require large amounts of energy, like experiments, demonstrations or events, like conferences or lessons. The Planner is not directly managing the operating point of devices, it rather urges the people working in REC to performance certain activities when it is more convenient for the EMS. As such, its presence or absence makes no difference to these experiments.

The devices present in the microgrid are grouped by nodes; for the Short-time Balancing System, a node will be the smallest participant. The balancing mechanism is initiated by either a node containing energy consumer or a node containing an uncontrollable source, whose operating point changes: a light gets switched on or off, or the output of a photovoltaic panel increases or decreases. When the operating point of a consumer increases, the effect on the system is the same as if the operating point of an uncontrollable source decreases: there will be a deficit of power (a negative imbalance), and additional power needs to be supplied. The reverse happens when the operating point of a consumer decreases or that of an uncontrollable source increases. At this point, the node that causes the imbalances signals to all other devices that an imbalance occurs and requests offers from devices to solve this imbalance. The only devices that are a possibly capable of dealing with the imbalance are the controllable sources, they answer the request for offers: for a negative imbalance each controllable source provides the amount of power it can supply and the cost, for a positive imbalance this will be the amount of

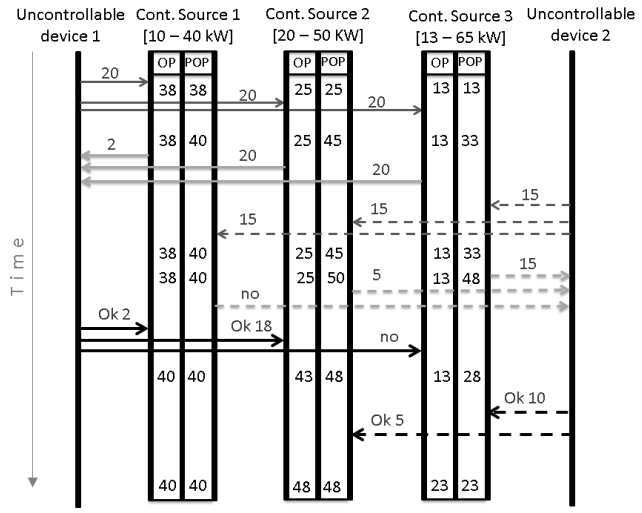


Fig. 1. An example of the balancing mechanism where the not-optimal distribution of power occurs.

power they can decrease and the profit. From this list of offers, the node that caused the imbalance chooses the option with the lowest cost or highest profit, potentially selecting multiple devices to cover the imbalance.

That way of choosing is a realization of a greedy algorithm for power production optimization, which should give the optimal solution when immediate communication between the agents is considered. However, in real experiments, when there are multiple devices that cause imbalances and multiple devices that can cover it, a sub-optimal behavior is observed. It is connected to the asynchronous operation of the multi-agent system and the delays in message sending between the agents. Consider the example in fig. 1: there are two consumers and three controllable sources. Their operating points are specified in the respective columns OP. At one point in time, an uncontrollable device 1 requests for an additional 20 kW. This can be covered by the two cheapest controllable sources. They report how much they can cover (2 kW and 20 kW respectively), and increase their provisional operating point (POP) - this is the operating point that they would have if the offer is accepted. Before the uncontrollable device accepts the offer, a second uncontrollable device causes an imbalance of 15 kW. As the provisional operating point of the cheapest controllable source is at maximum, it cannot send an offer. The second controllable source sends an offer for 5 kW, the third controllable source sends an offer for 15 kW. When the first uncontrollable device accepts the offers it got, this alters both the operating points and provisional operating points of the first two sources, this however does not change the offers that were sent to the second uncontrollable device. It accepts the offers it got, employing the more expensive source while leaving the second source still with a reserve of 2 kW. This is not the optimal situation; it can only occurs= when there are three or more controllable sources with different prices. In this case, it only occurs when a controllable device is near its maximum or minimum and an imbalance cannot be solved by one of the controllable sources alone. In reality, unlike in the example, the imbalances are of much smaller magnitudes so the non-optimality occurs at a smaller scale. Furthermore, future imbalances will again prefer to cheaper source, which also causes this sub-optimal situation to be quickly resolved, so this behavior is limited in time.

The standard behavior of the Short-time Balancing System is optimizing for the operation cost of the microgrid. While the Short-time Balancing System is capable of adjusting operating points quite optimally, it cannot decide on whether or not a controllable source needs to be powered on or can be powered off. This is due to the fact that the Short-time Balancing System does not consider predictions, but merely considers the current situation and reacts to immediate changes. The Planner, which should contain information on activities of energy consumers and prediction of the operating point of the uncontrollable sources, is capable of making such decision. This makes the Planner the first stage in the optimization process. However, the Short-time Balancing is a critical component in this process, and its operation can be verified and tested without the Planner.

The example of balancing behavior of Short-time Balancing System is presented in Fig. 2. In the setup, there were 5 consumers (aggregated to one for clarity of the figure), gas microturbine, engine, external power grid and two wind turbines. The time between recognition of imbalance to confirming the balancing action was by average 56 ms. It can be seen that the system is not using the external power grid unless there is really no other choice: when the operating point of a controllable source is near its upper or lower limit, the source has no more control capabilities in one direction. When this happens for all controllable sources, they no longer can maintain stable power in the microgrid without the help of the external network. The system is also reducing the operating point of the more expensive source, in this case the gas microturbine. The overall sum of production and consumption has to be maintained the same and the Short-time Balancing System is managing to achieve that.

IV. Calculating cost of operation

The cost of installation of renewable power sources, batteries and measuring system (smart metering) is still quite high, even though the popularity of such devices is growing. From an investor point of view, the decision of building such facilities has to be carefully calculated taking into account the future prices for electric energy, maintenance costs and the other factors. In this work, we are free from such dilemmas, as the REC is made to research the new technologies for power production and the reason for its installation is well defined and does not have to consider the return of investment time. That is also a reason why we would not consider amortization costs of operation of the research center, as it is included in the cost of fulfilling the goal of the building which is spreading and widening the knowledge about these technologies.

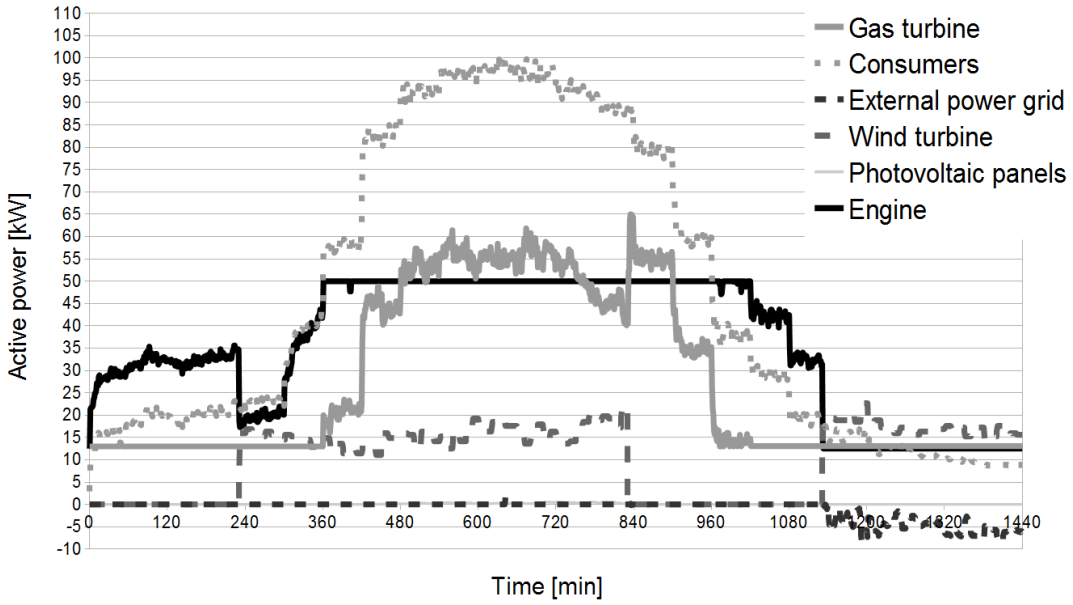


Fig. 2. The example operation of Short-time Balancing System.

The costs that are considered, are the cost of the current operation of the considered devices, under such conditions, the renewable power sources produce almost free energy. On the other hand the controllable power sources like gas micro-turbine and reciprocating engine need fuel to their operation and the cost of this should be considered.

In next paragraphs, the short description of the models of devices that were used in the system will be presented.

a) Gas microturbine: The gas microturbine modeled in this system has a nominal power of 65 kW with the allowed operating point between 20% and 100% of its nominal power [9]. This unit works as a cogeneration device (produces heat and electric power). The cost of producing electric energy by this device $K^{GM}(t)$ [PLN] can be defined by equation:

$$K^{GM}(t) = B^{GM}(t) W_p k_{jp} \tag{1}$$

Where: $B^{GM}(t)$ – usage of gas/biogas $[m^3/h]$, W_p– average calorific value $[kWh/m^3]$ estimated at 6.0 kWh/m^3, k_{jp}–cost of fuel unit [PLN/kWh] assumed to be 0.2 PLN/kWh.

The usage of gas can be calculated from the following equation:

$$B^{GM}(t) = \frac{P_e^{GM}(t)}{\eta_e^{GM}\left(\frac{P_e^{GM}(t)}{P_N^{GM}}\right) W_p} \tag{2}$$

where: $P_e^{GM}(t)$ – average electric power produced by gas microturbine during time t [kW], η_e^{GM}– electrical efficiency of gas microturbine [-], P_N^{GM}– nominal power of the source [kW].

The efficiency of gas microturbine depends on the operation point of the turbine and is described by the following equation:

$$\eta_e^{GM}\left(\frac{P_e^{GM}(t)}{P_N^{GM}}\right) = -0.196\left(\frac{P_e^{GM}(t)}{P_N^{GM}}\right)^2 + 0,419\left(\frac{P_e^{GM}(t)}{P_N^{GM}}\right) + 0.0387 \tag{3}$$

The change of efficiency depending on the operating point is presented in Fig. 3.

b) Reciprocating engine: The reciprocating engine is also powered by gas or biogas, its maximum power is 50 kW and it is also a cogeneration unit [9]. The cost of operation (without amortization costs) is expressed by the following equation:

$$K^{E}(t) = B^{E}(t) W_p k_{jp} \tag{4}$$

Where: $B^{E}(t)$ – usage of gas/biogas $[m^3/h]$, W_p– average calorific value $[kWh/m^3]$ estimated at 6.0 kWh/m^3, k_{jp}–cost of fuel unit [PLN/kWh] assumed to be 0.2 PLN/kWh.

The usage of gas can be calculated, similarly to the gas microturbine, from the following equation:

$$B^{E}(t) = \frac{P_e^{E}(t)}{\eta_e^{E}(\frac{P_e^{E}(t)}{P_N^{E}}) W_p} \tag{5}$$

where: $P_e^{E}(t)$ – average electric power produced by engine during time t [kW], η_e^{E}– electrical efficiency of engine [-], P_N^{E}– nominal power of the source [kW].

The electrical efficiency was estimated and is given by equation:

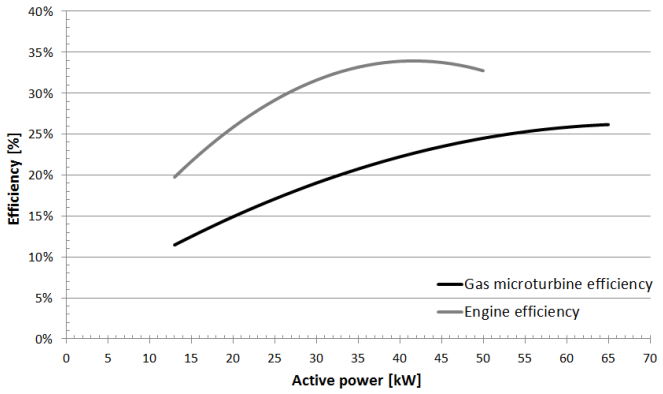


Fig. 3. The efficiency of producing power by the gas microturbine and reciprocating engine.

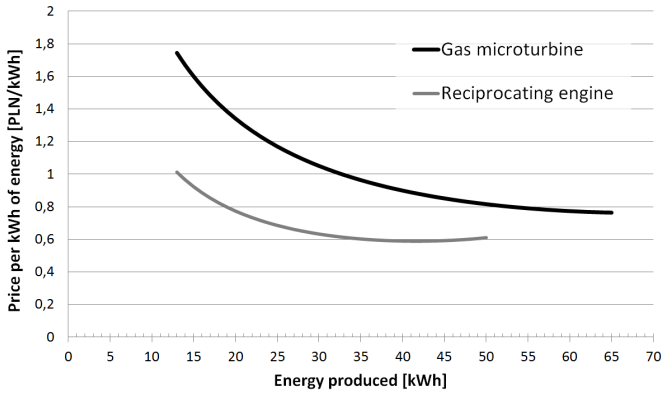


Fig. 4. The cost of producing fuel by the devices.

$$\eta_e^E\left(\frac{P_e^E(t)}{P_N^E}\right) = -0.432\left(\frac{P_e^E(t)}{P_N^E}\right)^2 + 0,72\left(\frac{P_e^E(t)}{P_N^E}\right) + 0.0395 \quad (6)$$

The diagram of electrical efficiency is also presented in Fig. 3. As can be seen the Engine has higher efficiency factor, so as a source for electrical energy it is more fuel efficient and therefor cheaper. However, when the cogeneration aspects are considered the gas microturbine might be more efficient.

The price of the controllable power sources dependent on the operating level are presented in Fig. 4.

c) External power grid: The current average price of power in the external power grid in Poland is fairly low [10]. The average cost per kWh is around 0.3 to 0.4 PLN, plus around the same value for the distribution of power and some additional fees. Comparing that to the price for energy produced by controllable sources, see Fig. 4, shows that the price of electricity from micrgrids own sources is higher. But it is not fully the case. Both gas microturbine and engine are cogeneration units which, in case of no connection to systematic heating system, are the most efficient heat sources. In cogeneration the efficiency of burning fuel is high and the cost of not using electricity for heating should be subtracted from the cost of producing electricity. What is more, in case of microgrids their own power sources have no additional fees for distribution of power or the maintenance of connection. Also, the internal production of energy is necessary in island mode operation. The combination of renewable and fuel-based controllable power sources allows to maintain fairly stable microgrid in case of blackout in the external power grid. Due to the fact that the renewable power sources are almost cost-free, the combination of renewable power source and gas microturbine gives the overall average prices that are much lower than any tariffs with the national power grids.

The assumption of the system was to keep the exchange of power between external power grid and microgrid minimal, but we did not have data to include full cost calculation on microgid side, that is why the prices of power from external power grid had to be set artificially. The cost of power from external power grid is set to 0.9 PLN/kWh for the energy taken from the grid (including distribution fee) whereas sending energy to the grid brings 0.05 PLN/kWh of profit.

d) Battery: The model of the microgrid includes battery and flywheel as a power storage units. The flywheel is a device that only works as a buffer of energy due to its inability to store the energy for long time. That is why it is not considered as a storage unit that can be managed by the system. The battery is a device that requires control at any point of its operation, in the simplest situation it needs the information whether it shall be in charging or discharging mode. The battery can be treated in EMS in two ways: or it is a device that takes part in the balancing, changing its operating point (setting charging or discharging mode) to supply power or consume power; or it is treated as intermediary device that is only a buffer between the producers and consumers.

The cost comparisons were done without the presence of a battery in either the agent based EMS or the optimal operating points calculator. The reason for this, is that there needs to be some prediction of the future power consumption and production so that the management system can decide whether or not it is a good time to charge the battery. This decision is impossible to make from just looking at the current power supply and demand situation. When the system has knowledge about the future energy supply and usage, ideally, the battery should be charged when there will be a surplus of energy to prepare it for discharging when there will be a deficit of power. Such knowledge can only come from a predictor that has a prior information about typical profiles. If the same predictor were to be used in either the agent based EMS or in the optimal operating points calculation, the battery will behave either as a normal consumer that consumes when there is a surplus of power (when the situation is good for charging it), or as the cheapest supplier when there is a shortage of power. In the end, this would not influence the cost calculation differently from adding a consumer or producer, but it would complicate the calculations and the visualization.

A battery is very useful in situation of variable prices in

external power grid, especially when the prices follow some averaged profile (e.g. represented by multiple tariffs, peak/off-peak prices). The storage device can include in its profiles and predictions the prices of power from external power grid and set its state (charging/discharging) to use the most of cheaper energy.

V. The calculation of optimal operating points

The first benchmark for multi-agent Short-time Balancing System is the system that simulates the optimal operating points for all controllable devices in the microgrid. We assume that it has information about the demand and supply of power in every moment in time. As such, it has enough knowledge to decide to switch on or off controllable power sources, which in some cases would be justifiable. This possibility was blocked to make it possible to compare the costs of operation with the operating points calculated by Short-time Balancing System. In the considered EMS the decision about switching on or off the controllable source is made by different module – the Planner that uses prediction of electricity and heating requirements to decide when and which devices should be active. The tests considered only Short-time Balancing System, and experiments were run without the Planner, thus without the possibility for switching off controllable sources. The reason for that is the same reason as why the experiments were run without the battery: both optimal and multi-agent system need to use the same predictor, which would only add more data without having an influence on the end result.

The algorithm for calculating the optimal operating points requires list of all active nodes and power sources, forecast of weather conditions and prediction of usage in each node. First, the program calculates the sum of consumption in given period, counting all active nodes. Then, it calculates the sum of production of power from renewable power sources in given period (sum from all photovoltaic panels and wind turbines). The remaining power is production and consumption that cannot be controlled; the difference between them is what really has to be balanced. With this knowledge, it is possible to calculate the operating point of controllable producers in a way to cover all imbalances. The idea is to first use to the maximum operating point the cheaper sources (excluding the power from external power grid) and then use the more costly. The remaining imbalances are assigned to the external power grid. Having all that data allows to calculate cost of production from controllable power sources, according to the equations presented in section IV. The cost of power from external power grid is calculated using fixed price per kWh and from it is subtracted the income from selling energy to the external power grid.

In Fig. 5 the example solution of the EMS system is presented. As can be seen, the system uses the cheapest energy first and uses the energy from external power when there is no other choice.

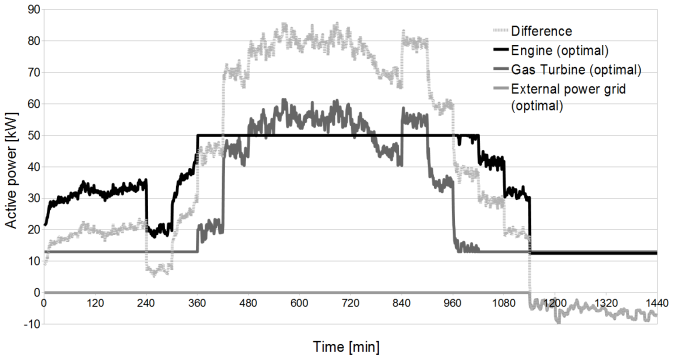


Fig. 5. The example of the operating point choice by the optimal algorithm with full knowledge.

VI. Managing with predefined profiles

The optimal operating point calculation has the advantage of knowing exactly the state of the microgrid in every moment in time. That is a perfect, but not possible situation. In practice the historical knowledge is used by defining the profiles of power usage and the profiles of weather conditions. It allows to set the operating points of controllable devices ahead, hoping that the prediction is good.

The consumption nodes are by default not controllable, in the system the power usage is a subject to planning, but it is an action that is executed by the people interacting with the system, not enforced by automatic control. The controllable power sources have to have some level of production set. We assume that there is some prior knowledge about the level of power usage in the microgrid, e.g. from past measurements. That allows to estimate the power needed to be produced to keep the power more or less in balance. The real balancing device will be the external power network that can provide or consume infinite amounts of power (actually it would be the amount limited by the connection that the microgrid has to the distribution network, but we make here a simplification). The renewable power sources operate with a goal to produce as much power as possible in the given conditions.

The calculator of cost for such scenario was developed, it is very similar to the algorithm of optimal operating points calculator, the main difference is calculation of operating point of controllable producers. In the calculation the same assumption as before is made that exchange of power between external power grid and microgrid should be minimized. The control assumes that the data about consumption and production from renewable power sources are given by the profile that has one averaged value per hour. The choice of the time interval is arbitrary, but it is a common time interval for profiles. That means that the operating points of controllable devices will be set once per hour. All the imbalances that appear are dealt with by external power grid.

The profiles are created by averaging the real consumption and renewable production levels, divide this into one-hour intervals and then setting the operating points for the controllable sources to this average for one hour. In the presented

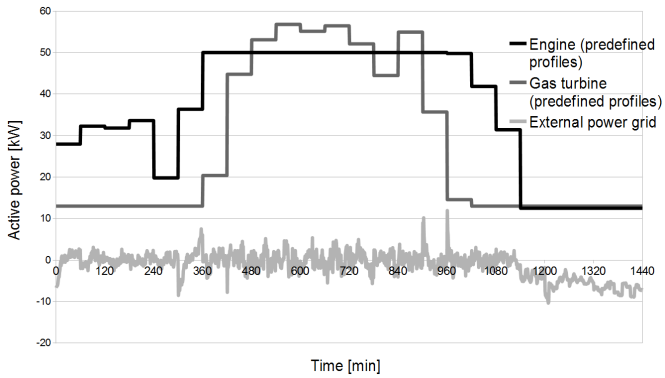


Fig. 6. The example of the operating point choice by using predefined profiles.

experiments, the averages of the simulated day are used; this makes the error of the profile relatively small. In a more realistic scenario, the averages would have been calculated using data from a longer period, yielding a profile that will exhibit a bigger error compared to the optimal. Because of the choice of the average profile over the same day, the amount of power produced by controllable sources during the entire day will match the amount of the optimal profile. Despite this, the averaging profile will perform worse, as it does not necessarily choose optimal sources or optimal operating points. The profile is therefore the best possible averaged profile for this day, which still provides an upper limit to the total cost.

In Fig. 6 the example solution of the microgrid management with predefined profiles is presented. As can be seen the system uses the power from external power grid more extensively, simply to maintain the balance.

VII. Cost comparison

The presented solutions have the same aim and work on the same data of weather conditions and face the same consumption profile. The solutions were compared and overlaid on a graph in Fig. 7. In the figure, only the controllable power sources are shown, as this is the only aspect where there can be a difference. The pairs of devices (engine and gas microturbine) from each solution are represented with the same color, but with different style of line. The solutions clearly follow the same pattern, which means all solution set the operating points not too differently from the optimal profile. The profiles for the controllable sources in the solution with predefined profiles deviates the most, as it uses the external network for balancing rather than the available controllable sources. The Short-time Balancing System has operating points that very close resemble the profiles of the optimal solution. It has some small delays in changing the operating point, but that is due to the fact that it is the only EMS that was simulated in real time: some time is needed to detect an imbalance. Different simulations show the same and expected behavior, as was described earlier.

At 240 minutes, it seems Short-time Balancing System changes the operating point earlier than the other solutions.

TABLE I
Comparison of costs using Short-time Balancing System, optimal calculation of operating points and the EMS with predefined profiles.

	Short-time Balancing EMS	Optimal solution	EMS with predefined profiles
cost of gas	1276.79	1276.54	1276.295
cost of power from external power grid	-2.369	-2.838	6.670
total cost	1274.421	1273.703	1282.959

This is a side effect of aggregating the data in time: the devices set or report their operating point once per minute, but this interval is not synchronized between devices. Aggregating the data to match the time scale of the other solutions, to be able to show it on a graph, can yield the side-effect as it appears at the 240 minute mark.

The cost of operation was calculated, using the equations presented in section IV. The result of cost comparison is presented in table I. It can be seen the Short-time Balancing System does not differ much from the optimal solution. The cost of fuel (in this case gas) is very similar in all three approaches, because, as can be seen in Fig. 7, the total fuel needed for all controllable sources during the day are very similar. The operating point of system with predefined profiles is an average of the optimal profile, so its gas usage is very similar. Because EMS with predefined profiles uses the external power grid to balance its cost of using it is visibly higher.

The Short-time Balancing System is suffering from delays in first detecting the change and then reacting to it. These delays always occur, both with increase and decrease of power, consequently over long time they partially even out, drawing the cost closer to the cost of optimal solution.

The actual cost of operation of the microgrid depends on many factors, the prices of gas and power from external power grid are fixed, which in realistic conditions could be changing or simply different. The important here is that the distributed system, which thanks to its agent architecture is more robust to failures, can set operating points to almost optimal values.

VIII. Conclusion

The Short-time Balancing System is a distributed EMS that is successfully managing to balance the power in the microgrid within required time (the average is below 60 ms for average balancing time, on an accelerated simulation using a single computer). The way it is adjusting the operating points of controllable devices is following the pattern of the optimal solution. The cost comparison without real system is difficult to define as some simplifications have to be done, e.g. not allowing the systems to switch off the power sources when they are not needed. The presented comparison here was aimed at comparing the performance of the system with the optimal setting of operating points and with a system that only knows the predefined profiles of consumption and renewable source

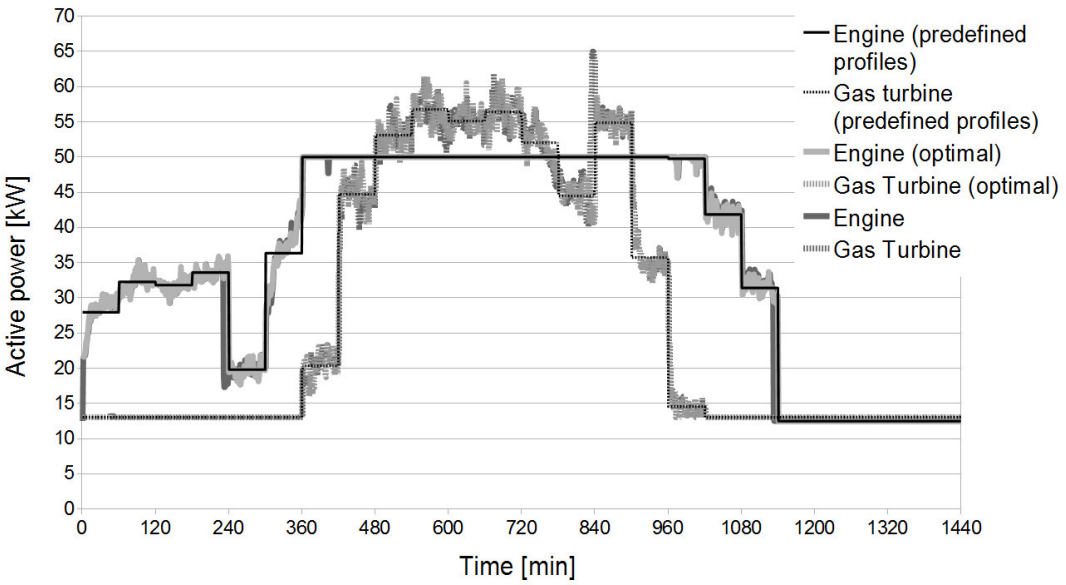


Fig. 7. The comparison of operation of controllable power sources and external power grid using different solutions.

production. The most basic comparison involves the operating points, which directly define the cost of fuel usage, which consequently is a good comparison factor. To fully evaluate real costs of operation, the cogeneration units have to evaluated considering the cost and usage of heating at the moment of tests, for example cost of microturbine and engine should be corrected by the factor of not used amount of power for heating the buildings in REC.

The way the microgrid uses its own production sources depends on the choice of the aims of the EMS system. In the example used in this article, external network and local controllable sources were used, and conclusions from this data were drawn. This is merely to illustrate the capabilities of the Short-time Balancing system. Under the assumptions made, if the main goal is to operate microgrid as cheaply as possible there will be a much larger usage of power from the external power grid than in situation where the independence from the national power grid is rewarded. Different settings in the cost functions of different devices allow for changing the behavior of the balancing strategy, additional sources will be selected according to their cost profiles.

Comparison shows that the Short-time Balancing System is working almost optimally. It is reflected by the cost of operation which is also very similar.

ACKNOWLEDGMENT

The research of W. Radziszewska was supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing. Project financed from The European Union within the Innovative Economy Operational Programme 2007-2013 and European Regional Development Fund.

REFERENCES

[1] R. Lasseter, A. Akhil, C. Marnay, J. Stephens, J. Dagle, R. Guttromson, A. S. Meliopoulous, R. Yinger, and J. Eto, "White paper on integration of distributed energy resources: The certs microgrid concept," CERTS, Tech. Rep., April 2002.

[2] A. Tsikalakis and N. Hatziargyriou, "Centralized control for optimizing microgrids operation," *Energy Conversion, IEEE Transactions on*, vol. 23, no. 1, pp. 241–248, March 2008.

[3] M. D. Ilic and S. Liu, *Hierarchical power systems control : its value in a changing industry*, ser. Advances in industrial control. London, Berlin, Paris: Springer, 1996. [Online]. Available: http://opac.inria.fr/record=b1104385

[4] M. Vasirani and S. Ossowski, "A collaborative model for participatory load management in the smart grid," in *Proc. 1st Intnl. Conf. on Agreement Technologies*. CEUR, 2012, pp. 57–70.

[5] W. Radziszewska, Z. Nahorski, M. Parol, and P. Pałka, "Intelligent computations in an agent-based prosumer-type electric microgrid control system," in *Issues and Challenges of Intelligent Systems and Computational Intelligence*, ser. Studies in Computational Intelligence, L. T. Kóczy, C. R. Pozna, and J. Kacprzyk, Eds. Springer International Publishing, 2014, vol. 530, pp. 293–312. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-03206-1_20

[6] P. Pałka, W. Radziszewska, and Z. Nahorski, "Balancing electric power in a microgrid via programmable agents auctions," *Control and Cybernetics*, vol. 4, no. 41, pp. 777–797, 2012.

[7] J. Wasilewski, M. Parol, T. Wojtowicz, and Z. Nahorski, "A microgrid structure supplying a research and education centre - Polish case," in *Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on*, 2012, pp. 1–8.

[8] W. Radziszewska and Z. Nahorski, "A multiagent energy management system for a small microgrid equipped with power sources and energy storage units," in *International Congress on Energy Efficiency and Energy Related Materials (ENEFM2013)*, ser. Springer Proceedings in Physics, A. Y. Oral, Z. B. Bahsi, and M. Ozer, Eds. Springer International Publishing, 2014, vol. 155, pp. 411–417. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-05521-3_53

[9] D. Kowalska, M. Parol, J. Wasilewski, and T. Wójtowicz, "Opracowanie uproszczonego (przez przyjęcie zintegrowanych modeli urządzeń i instalacji cieplnych) projektu sieci (fragmentów instalacji) cieplnej w poszczególnych obiektach ośrodka badawczo-szkoleniowego, łącznie z określeniem odbiorów ciepła oraz doborem rozproszonych źródeł ciepła (kolektorów słonecznych, kotłów do spalania biomasy), a także zasobników ciepła (urządzeń grzewczych, urządzeń chłodniczych)," Systems Research Institute PAS, Tech. Rep., 2010.

[10] ARE s.a, "Energia.pl," http://www.energia.pl/porownanie_obrot.php.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 281–288
DOI: 10.15439/2015F402
ACSIS, Vol. 6

The Architecture of an Information System for the Management of Hybrid Energy Grids

Olha Shulyma
Sumy State University, 2, Rymskogo-Korsakova st., 40007 Sumy, Ukraine;
Malmö University, 205 06 Malmö, Sweden
Email: o.shulym@opm.sumdu.edu.ua

Paul Davidsson
Internet of Things and People Research Center, Malmö University, 205 06 Malmö, Sweden
Email: paul.davidsson@mah.se

Vira Shendryk, Anna Marchenko
Sumy State University, 2, Rymskogo-Korsakova st., 40007 Sumy, Ukraine
Email: {ve-shen, nenja_av}@opm.sumdu.edu.ua

***Abstract* — The objective of this paper is to present the design of an information system using software agents. The features of the decision-making process concerning a working model of a hybrid energy grid and the functional requirements of the information system are described. Based on an analysis of the information flow occuring during the data processing for making appropriate decisions, a system architecture which combines a set of independent platforms, is proposed. Multi Agent System is used for the implementation and integration with other systems, such as a power system analysis tool, as well as presentation and storage tools.**

I. Introduction

From a production perspective, electrical energy production has become more difficult as traditional big power plants use fossil fuel and therefore are regarded environmentally inappropriate, and as nuclear power plants are regarded inappropriate due to safety. However, to generate energy in large conventional power plant is cheap and not technically difficult. Furthermore, energy distribution to the user could be very expensive due to the significant losses that occur. Many countries are presently introducing the concept of distributed energy production and consumption by using renewable energy sources (RES). This means that the power system mainly consists of medium and small generators that are located directly next to the end user [13] The user cares individually about the use and maintenance of RES. Electricity is generated for user's own needs, and if there is excess, the user can sell it to the grid where other users can use this energy. The presence of several sources producing electricity, make such a system a hybrid system [1]. Successful examples of using solar and wind energy have already been shown, and the construction of hybrid energy systems has become more attractive.

In addition, RESs are becoming increasingly popular as they produce electricity in a more environmentally friendly way, and therefore are getting financial support from governments. In developing countries often the idea of using RES comes from private stakeholders. However, it is difficult for laymen to operate such grids with maximum benefit as the system is characterized by rapidly changing operating modes and configurations are depending on external conditions.

The decision making process concerning the management of a distributed grid is complicated by the presence of a large amount of discrete variables that are affecting the system. It can take a long time to make a decision as you want to be sure that the right decision was made with respect to maximize benefit. Thus, it may be convenient to use Decision Support Systems (DSS), which can take on some the user's tasks. However, before the creation of a DSS, it is necessary to build an information model that can be used to provide the appropriate information to the DSS. It is necessary to determine exactly, what type of tasks that will be taken over by the system, how to solve these tasks, and how to provide the necessary data to the system.

The key goal of this research relates to the development of the architecture of an information system for assessing the current condition of the grid and predicting the future behavior of the grid

To achieve this goal we have identified a set of research tasks, which have their own sub-goal:

- Analyze the state-of-the-art in creating DSS by doing a comparison of existing approaches with the aim to determine their usefulness, strengths and weaknesses.
- Determine the features of the study object - hybrid energy system. Here we should consider not only the physical structure of system, but also features such as weather conditions, users' behavior, etc., which influence the system.
- Determine what information should be provided by the DSS, which can also be regarded as requirements for system functionality.
- Determine the whole cycle of the process of making decision concerning developing information model. At this stage, it should take into account not only the results that will be obtained in the previous stages, but also the ways of software implementation.
- Describe the architecture of the system

II. Literature Review

There is a lot of research studying the operation of the grid. In this section, we present a review of the previous work related to the planning of grid with RES.

A vast majority of the systems have been designed to work with hybrid energy system in a particular region. During the planning process of such power grids, two types can be distinguished:

- Stand-alone hybrid system [8, 9, 14]
- System with connection to an external grid [4].

Also, the DSS differ in their ultimate goal:

- Prediction system to determine the type of the used resource [10].
- System for investigating the performance based on the optimum design of hybrid systems [8, 6, 14, 10]
- Prediction system to determine the output and the consumption of energy [2, 7].

Moreover, the DSS differ in their modeling approach:

- Systems based on simulators for dynamic models [9, 16, 4, 10].
- Systems with analytical models [2, 8, 6].

Finally, for making decisions, special frameworks can be used [4, 10] or systems can combine special techniques to make decisions: linear programming dynamic programming, multi-objective methods, etc.

Brief descriptions of the systems under consideration are given in Table 1.

Although extensive research has been done, it mainly focuses on determining the optimum number of PV/wind power generators. On the other hand, some models are proposed to predict the performance of an energy system, but do not address the energy management for scheduling consumption and selling energy. Thus, it would be useful to build a universal system that could solve all these tasks. In addition, all systems were constructed integrally with the minimum number of independent modules and most of them do not offer any interactivity. Especially complex and confusing for the layman is the treatment of the grid model. Thus, the DSS architecture we propose should take into account all these factors, and based on them and also the physical structure of the grid.

Summarizing all the above, it was decided to develop an information model that is not for a specific grid that physically exists, but a more general model. This decision is based on the desire to build a common information system that can be applied to any laymen in the planning stages of building a grid, as well as for further analysis of its operation. This abstract model also requires the definition of a possible physical implementation, featuring the functionality of its constituent parts.

The DSS architecture proposed in this paper is taking into account all these factors.

Table I.
Brief Description of Systems

Reference	System Description
[2]	A method for calculating the optimum size of hybrid photovoltaic (PV)/wind energy system, with performance by hourly basis.
[10]	A technical and economic study to design a hybrid PV/wind power generating system for one domestic in India.
[6]	A technique for analyzing the performance of the autonomous PV/wind hybrid energy systems.
[9]	A technique to determine the optimum size of hybrid PV/wind energy system.
[14]	A computer-based approach for evaluating the general performance of stand-alone PV/wind generating systems.
[8]	A new integrated tool and a framework for making complex decision process in the UK energy sector.
[4]	A new framework that includes tools for the monitoring and management of housing energy systems. The proposed tools are tested on realistic instances based on the Italian electricity market

III. The Features of the Considered Energy System

The term hybrid renewable energy system (HRES) is used to describe any electric power system with more than one type of renewable energy source such as PV, wind, and PV/wind. [7] Such system can be also connected to an external grid.

The type of HRES considered in this paper is illustrated in figure 1. There is a group of buildings with installed PV (solar) panels. There is also a common park of wind generators and an energy storage bank. Furthermore, there is a connection between the hybrid energy system and an external grid, both for covering consumption peaks and for selling surplus energy.

In every house there is a work schedule of electrical equipment, as well as a model of additional electricity consumption, depending on the daily routines (weekdays, weekends and holidays). Note that in the schedule, the

different working days may vary, power consumption may differ significantly according to this schedule. For instance, the consumption at night and holidays may differ from the daily working at 10-12 times.

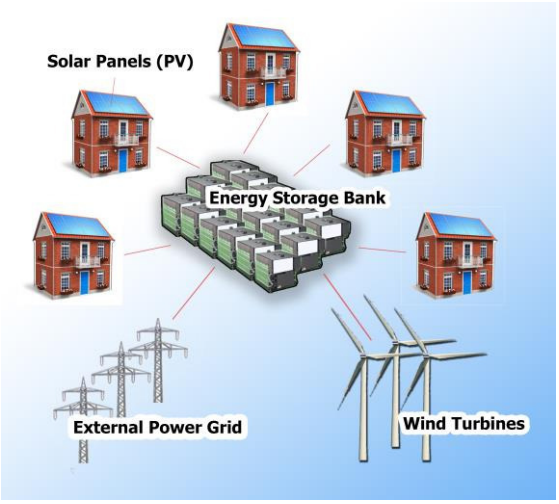


Fig. 1. A Schematic Idea of the HRES studied

The question of which renewable energy source to choose has been considered in [16]. The reason for using a common park of wind turbines is that in order to generate a large amount of electricity, wind turbines should be placed on a flat ground with no barriers to air flow. Installation of small wind turbines on the roof of each building does not give enough power in a temperate continental climate. In addition, it should be noted that in most areas with a temperate climate, the of wind intensity changes throughout the year. In the winter, the wind is much stronger than in the summer. For the wind generator setup, this means that in the summer the produced energy is 1.5–2 times less compared to the winter [15].

The intensity of solar energy also is very different depending on the season. If the amount of solar energy on a summer day is 5–7 kWh/m2, it may fall to 0.8–1.2 kWh/m2 on a winter day. Thus, wind turbines work best in winter, and solar panels in the summer. Therefore, it is logical to combine these energy sources into one system. The challenge is to choose power of wind turbine and solar panels in such way that energy production meets the consumption demands during the whole year.

The existence of several sources generating capacity in the system, as well as the mechanisms for storage and sale of electricity, causes the need of a large amount of input data for the construction of an appropriate information system for forecasting energy consumption and sales. Moreover, there is a large number of output and other information that the system should provide. Thus, before designing the system, it is necessary to determine what questions that must be answered by the system and based on them to determine the necessary input information.

IV. Questions to the Functioning of the Information System

The structure of hybrid grid was determined in previous sections, and in previous research [17] was given the analysis of existing methods, that can be applied in information system and tasks of the information simulation of distributed power grid. Thus, on this step can be determined functional requirements, which describe functions that the system executes. They are sometimes known as capabilities [3].

The user of the system eventually wants to know when, from what resources and what amount of energy can be produced or consumed, as well as the best time of the power storage and sale. Thus, were determined next groups of functions with their sub-functions, which helps to get answers:

- Gathering information on:
  - weather forecasts for the area (temperature, solar insolation, wind speed, cloudiness, etc.);
  - rates of "green tariff";
  - technical characteristics of the object of observation (the capacity of the grid, the amount of solar insolation etc.);
  - the energy consumption by users.
- Working with data:
  - to transmit collected data to the system server;
  - to perform data processing and storage;
- Analytic functions:
  - Hourly forecast of the production of electricity from renewable energy sources, based on the forecasting the weather.
  - Prediction of energy consumption based on appropriate forecast, which operates hourly historical data about energy consumption during the day in each month of the year;
  - Prediction of electricity sales, based on appropriate forecast, which operates data about “green tariffs”.
- Providing the visual representation of information in graphs, tables, and reports.

After determining the functionality, it becomes possible to present the information model of the system. Modelling enables a common and comprehensive understanding of a process and can serve as a basis for developing the architecture.

A Data Flow Diagram (DFD) shows the Information View as required data flows between the functions of a system. A summarized scheme of the decision making process at the top level is shown in Figure 2.

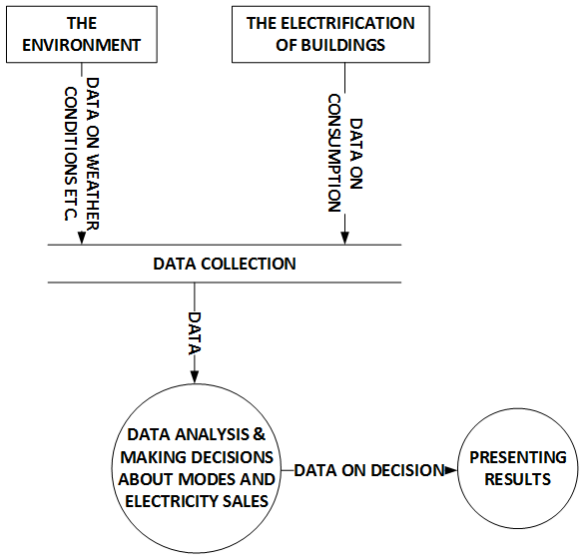


Fig. 2. The Information View of Decision Making Process

Both "The Environment" and "The Electrification of Buildings" transmit data that corresponds to the displayed list of functional requirements on "gathering information". "Data Collection" provides storage, conversion and transmission of the received data. "Data Analysis & Making Decision" performs analytic functions and transmits the output data for reports presentation to a user on the production, consumption, and sale of electricity.

To display all of the functionality of the system in accordance with this process, it is necessary to define the basic subsystems and functions to divide between them.

V. The Distribution of Functions Between Components of Information System

The main actions that the information system should provide are the collection, transmission and storage of information, data analysis, modeling and forecasting, the optimization of consumption and providing recommendations for management decisions, and also providing results in a convenient and understandable form for the users. This makes the structure of the system and it is necessesary to divide it into to several interacting subsystems.

As there are a great variety of approaches of implementing the operation of such systems, the individual subsystems could be implemented in different ways, and we can select the one most suitable for the task. Furthermore, this approach provides an opportunity to improve one part of the system without affecting the others when new requirements on the subsystems or technologies emerge.

Table II summarizes how the different subsystems are implemented in the first version of the system.

TABLE II.
THE BRIEF DESCRIPTION OF SUBSYSTEMS

Subsystem	The Way of Implementing
The Modeling Grid Work	The construction simulation models in Simulink.
The Data Storage	Database development with the use of MySQL Server.
The Data Collection and Transmission	Collection and transfer data to a database server and their processing.
The Formation of Recommendations about Grid Work and Sales Mode	The use of methods of forecasting and optimization in Matlab, when working with a simulation model based on the data collected. Methods are used to achieve the maximum benefit when consuming and selling electricity.
Results Presentation	Web-server and GIS-server to provide information to the user in a clear and simple form with the view of a particular grid on the map.

In addition, a web interface is used for user input of the desired characteristics of the grid (about the usage of PV panels and/or wind turbines, energy consumption mode, etc.). The reason for presenting information in this form is that the modeling of the grid and decision-making are provided by mathematical software packages with interfaces that are complex to understand for the layman.

In this paper, decision-making is based on modeling of power supply system with the physical features which have been identified above. Initially, it was told that the system should be universal, but there are different grids that connect the system and the system itself may consist of a number of elements and different capacities.

The solution of this problem is provided by the creation of a interface where the user can specify the desired scheme of the grid. But this raises the question of how to model and analyze a system that is not fully known. We propose the use of software agents to handle this.

A multi-agent system is a combination of several agents working in cooperation and to meet certain challenges in order to achieve the system goals. Agents are autonomous in that they can operate without human intervention, are socially-oriented in how they interact with other agents through the language of communication that is understood by all participants. Agents can also sense and respond to environmental perturbations. A multi-agent system is active and able to take initiatives to cause behavioral changes in the grid.

An agent-based approach will help maintain the positive features of the classic client-server approach and overcome most of its weaknesses. When using the agent-based approach, the application is logically divided into a plurality of components having a large degree of autonomy and are

able to communicate with each other. Also, the use of agents makes it possible to balance the load between multiple computing subsystems and conduct parallel processing of data by the asynchronous communication between the agents and the possibility of moving the agent code on the network. Application of this method allows to make full use of existing computing resources and thus solve the problem more efficiently.

There are many approaches based on the principle of assigning an agent for each grid element. This makes it possible to organize the work efficiently, since in this case, each agent is responsible only for a grid element. The agent contains the parameters and characteristics and performs calculations associated only with this element. We propose to use agents not only for the grid simulation model, but also to provide all the functions in the system.

Agents can be built according to the specification FIPA (Foundations for Intelligent Physical Agents). According to its agent platform via TCP/IP connected to external platforms [18]. To determine the set of agents necessary to distribute the system functions among its elements, we analyzed the information flows in the system, which are presented in Fig. 3. In the figure all processes for reading and transmitting data are operated by agents.

During the analysis of information flows, we arrived at the set of agents and their interaction that is shown in Fig. 4.

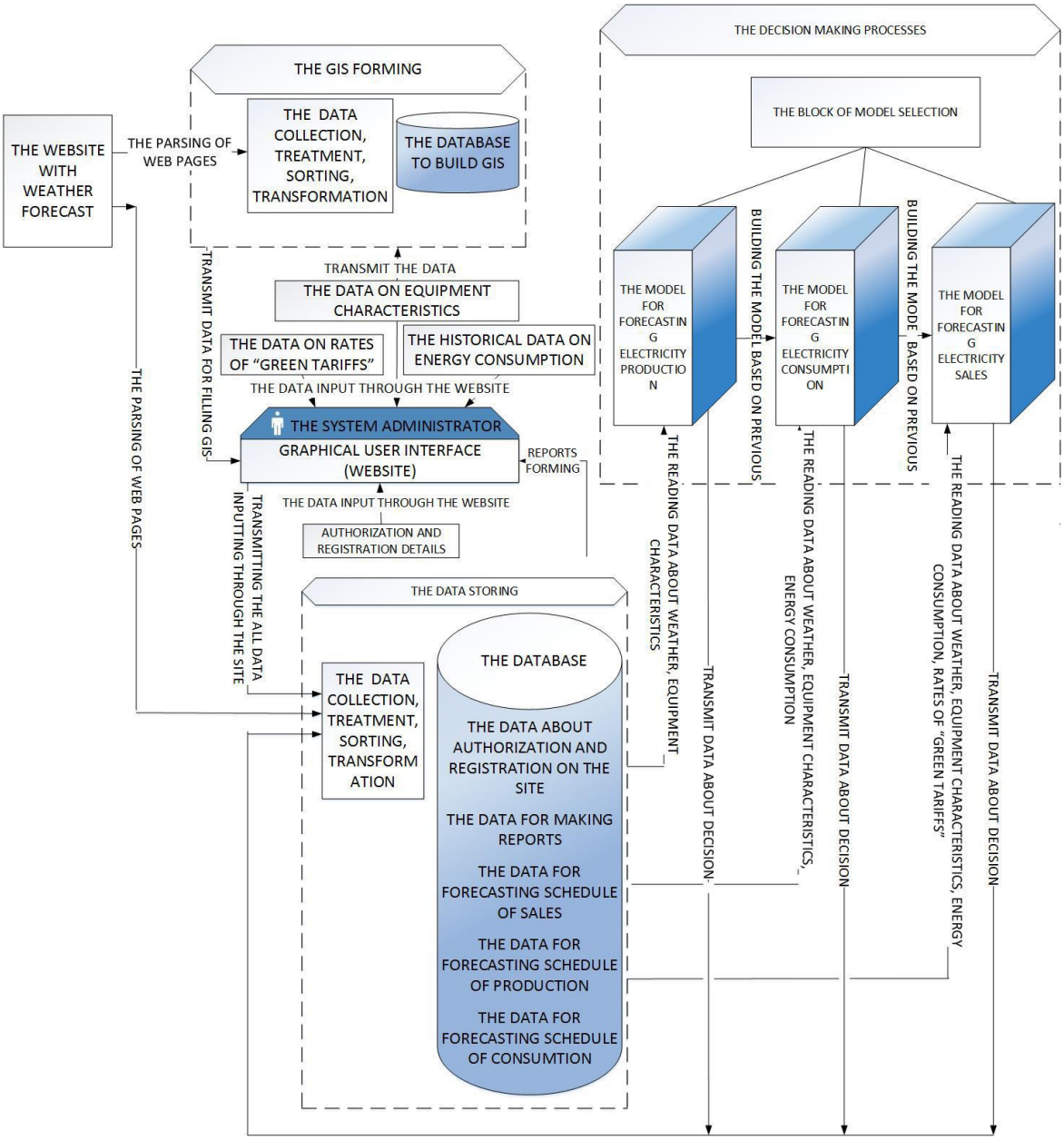


Fig. 3. The Information Flows in System

In the analysis and design of agent-based systems methodologies are used. In this case, as the methodology was used Gaia. According to it, the projected system is represented as an organization, which functions by implementing a set of roles. In the context of the Gaia Agent is an active program module which plays a set of roles.

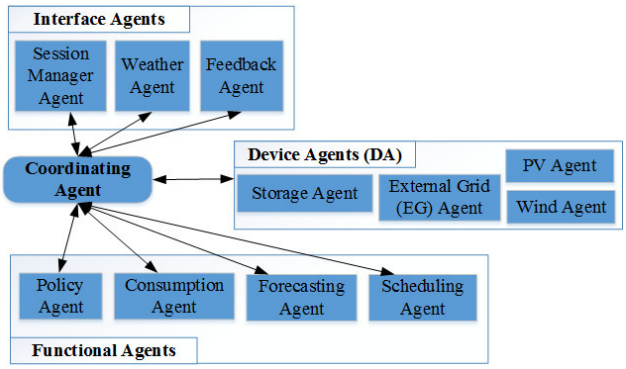


Fig. 4. An interaction between agents

Agents are grouped by their functional purpose:

- Interface agents - provide interaction with the user, e.g. presenting output information to the user in a convenient form.
- Device agents - gather information on the operation of technical objects in the hybrid power system (e.g., control switch on/off).
- Functional agents - transmit relevant data for modeling, forecasting and optimization: load, generation and price predictions, for optimal using RES for generations and sales.
- Coordinating agent - receives all partial or local plans from individual agents, analyses them to identify potential inconsistencies and conflicting interactions.

VI. THE ARCHITECTURE OF THE INFORMATION SYSTEM

At present, the term “architecture” is used very widely in the practice of designing systems, but it has many different interpretations. In accordance with the international standard for architecture descriptions of systems and software ISO/IEC/IEEE 42010 Systems and software engineering - Architecture description [11], the system has an architecture that can be described from various views of stakeholder, by considering the system architecture. Each point of view on the architecture of the system corresponds to viewpoint, which is based on a set of models kind. In [5] authors determined two points of view:

- A conceptual viewpoint, which describes the system in terms of major design elements and their interactions;
- An implementation viewpoint, which gives a view of the system in terms of modules or packages and layers;

An architecture description of the system from an implementation viewpoint, including agents, but without describing all processes is presented in a figure 5. It should be mentioned, that an information system has a three-tier architecture using software agents. At the first level architecture clients are presented and through created for them an interface they can access to the system and obtain information in a simple to understand form (graphs, charts, maps). On the second level is the logic implemented (charts, models and forecasts), and on the third, storage (the database level).

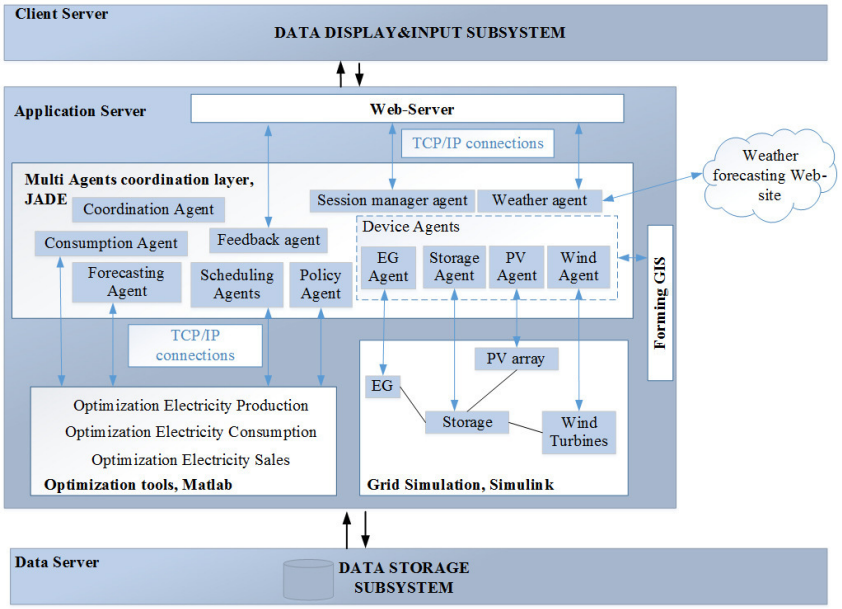


Fig. 5. The architecture description of the system from an implementation viewpoint

This information & analytical system of the electrification of buildings and forecasting electricity sales is planned for implementation as a web resource for authorized access of users are providing resource management and operational analysis of the management system of electrification. Thus, before creating the system, it has to divide the functions and actions of system users.

VII. The Use Case Diagram of The Information System

For determination of all possible expected behavior of the system, Use Case Diagram are often used. They show a set of use cases and actors (a special kind of class) and their relationships [12].

Use Case Diagram combines actors and use cases, and the relationships between them define a small set of relationships to structure actors and use cases. Some use cases relate to others by:

- Extending - An extend relationship implies that a Use Case may extend the behavior described in another Use Case, ruled by a condition.
- Including - An include relationship means that a Use Case includes the behavior described in another Use Case.

Generalization relationships in this case are presented between actors and means that “child” inherits all features and associations of the “parent”, and may add new features and associations.

The main functions of the system’s users are presented in Fig. 6 in the form of a Use Case Diagram. The use cases can be seen as a set of requirements to the system or responses that end-users of the system would like to receive. In addition, this diagram allows to determine the users and boundaries of the system, as well as a possible system interface.

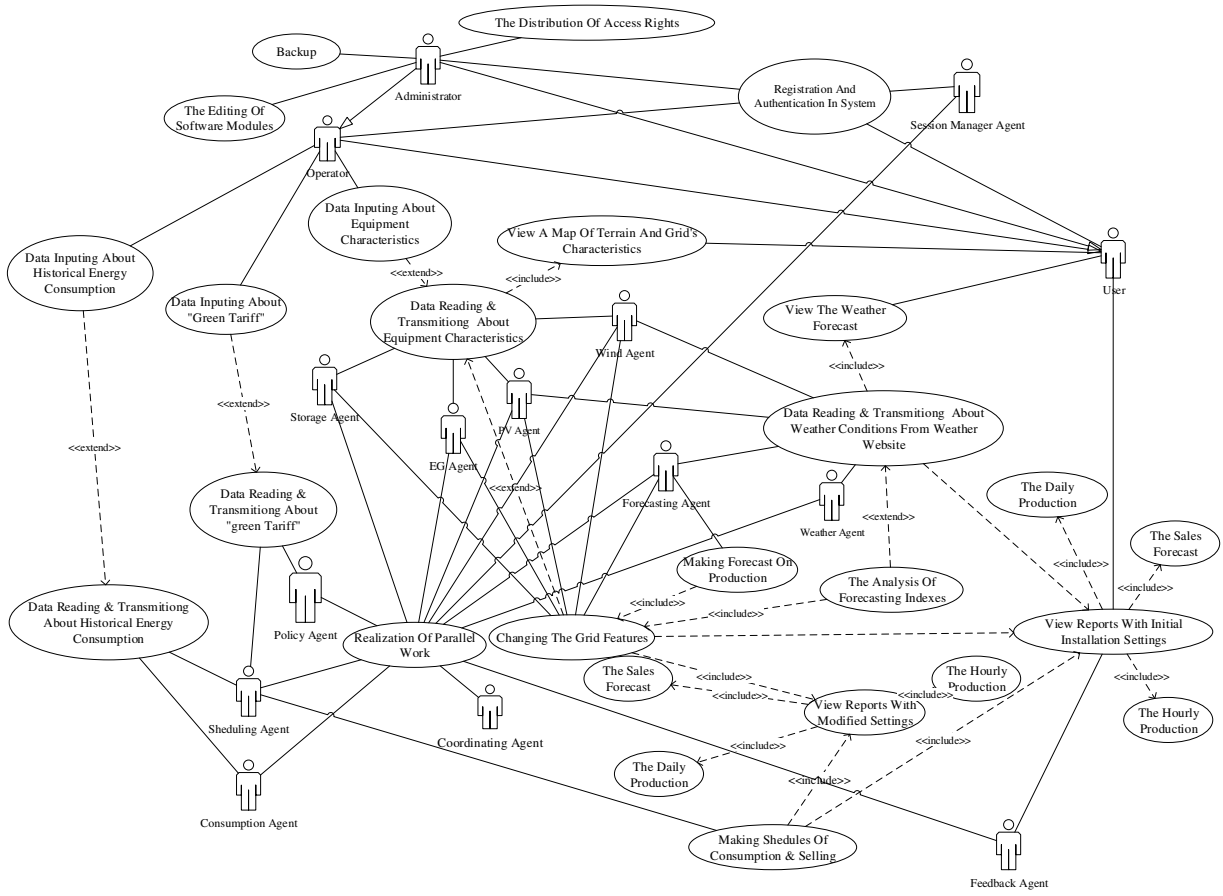


Fig. 6. The Use Case Diagram of the System

The actor “Administrator” should have full rights of web-resources management. The Administrator is responsible for the allocation of access rights to the users of the system, has the right to make and edit all the data for analysis, and decision-making and adjustment of all output data. Also the administrator is responsible for content filling inside the website, database administration, and creation of backups.

The functions of the "Operator" are data collection of information resources and entering the collected data into a unified database. During entering all data are verified and after successful verification record into database. A person enters these data, as it cannot be derived automatically from external resources.

The user and the owner of the power grid act as one entity "User" that performs analysis of the recommendations made by the system on a daily basis.

The actor "Website with weather data" is an external system and its functionality is automated completely.

VIII. The Further Development

The set of simulation models will be the basis of the decision-making process, and processing of information flow in the system is considered by combining modeling based on agents with simulation using dynamic models.

The development of software implementation of an information system as website has begun according to the architecture described above. In creating the system, the following main components can be distinguished: Web Server; Website; Agents Platform; Decision Maker Server based on simulation modeling; Database Server; GIS Server.

For their development we are considering the following tools: Server Platform NodeJS; a combination of HTML5, CSS and JavaScript for website, creating; JADE for implementing agents; MatLab to create a simulation model and decision-making; MySQL Server 5.5 as the database server; Quantum GIS as a server for the creation geographical information subsystem.

IX. Conclusions

Making decisions about the operation of hybrid distributed energy grids is a decentralized problem that can be addressed through building appropriate information and analytical DSS systems. In accordance with the presence of large amount of scattered information and the different tasks involved in the operation of the system, it is appropriate to use in a number of different components in the system. Therefore, we proposed a multi agent system platform for the implementation and integrating with others, such as power system analysis, presentation, and storage tools.

This paper addressed the problem of a scientific substantiation and solving problems with efficiency in data management for information flows in distributed hybrid energy grids through the development of information models of the system. Based on the analysis of previous research we proposed an information system using agent technology and provided an architecture description of information system for the prediction of hybrid power generation and energy sales. It was built according to the analysis of information flows that occurs during the processing data for making appropriate decision and based on set of developing information flow.

The whole system was presented in the form of parts according to the three-tier architecture. We also identified agents corresponding to each process going on in the system. We believe that proposed architecture and information system for the management of renewable energy sources and further implementation will allow to efficiently carry out the management of grid productions, sales and support.

References

[1] Ackermann, T., Andersson, G., & Söder, L. (2001). Distributed generation: a definition. Electric power systems research, 57(3), 195-204. doi:10.1016/S0378-7796(01)00101-8

[2] Ai, B., Yang, H., Shen, H., & Liao, X. (2003). Computer-aided design of PV/wind hybrid system. Renewable energy, 28(10), 1491-1512. doi:10.1016/S0960-1481(03)00011-9

[3] Abran, A., Bourque, P., Dupuis, R., & Moore, J. W. (2001). Guide to the software engineering body of knowledge-SWEBOK. IEEE Press.

[4] Barbato, A., Capone, A., Carello, G., Delfanti, M., Falabretti, D., & Merlo, M. (2014). A framework for home energy management and its experimental validation. Energy Efficiency, 7(6), 1013-1052. doi: 10.1007/s12053-014-9269-3

[5] Bass, L., Clements, P., & Kazman, R. (2013). Software architecture in practice. Upper Saddle River, NJ: Addison-Wesley.

[6] Celik, A. N. (2002). Optimisation and techno-economic analysis of autonomous photovoltaic–wind hybrid energy systems in comparison to single photovoltaic and wind systems. Energy Conversion and Management, 43(18), 2453-2468. doi:10.1016/S0196-8904(01)00198-4

[7] Deshmukh, M. K., & Deshmukh, S. S. (2008). Modeling of hybrid renewable energy systems. Renewable and Sustainable Energy Reviews, 12(1), 235-249. doi:10.1016/j.rser.2006.07.011

[8] Elhadidy, M. A. (2002). Performance evaluation of hybrid (wind/solar/diesel) power systems. Renewable Energy, 26(3), 401-413. doi:10.1016/S0960-1481(01)00139-2

[9] Gomaa, S., Seoud, A. A., & Kheiralla, H. N. (1995). Design and analysis of photovoltaic and wind energy hybrid systems in Alexandria, Egypt. Renewable energy, 6(5), 643-647. doi:10.1016/0960-1481(95)00044-K

[10] Hunt, J. D., Bañares-Alcántara, R., & Hanbury, D. (2013). A new integrated tool for complex decision making: Application to the UK energy sector. Decision Support Systems, 54(3), 1427-1441. doi:10.1016/j.dss.2012.12.010

[11] ISO/IEC/IEEE Std 42010:2011 – Systems and software engineering – Architecture description. Los Alamitos,CA: IEEE, 2011.

[12] Jacobson, I., Rumbaugh, J., & Booch, G. (1999). The unified modeling language user guide. Addison Wesley.

[13] Multin, M., Allerding, F., & Schmeck, H. (2012, January). Integration of electric vehicles in smart homes-an ict-based solution for v2g scenarios. In Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES (pp. 1-8). IEEE. doi: 10.1109/ISGT.2012.6175624

[14] Nehrir, M. H., LaMeres, B. J., Venkataramanan, G., Gerez, V., & Alvarado, L. A. (2000). An approach to evaluate the general performance of stand-alone wind/photovoltaic generating systems. Energy Conversion, IEEE Transactions on, 15(4), 433-439. doi: 10.1109/60.900505

[15] Solovjova, E. G., & Kondratenkov, A. N. (2013). Sistema avtonomnogo energo-snabzhenija zdanija v uslovijah II klimaticheskoj zony.

[16] Shendryk, V. V., Vashhenko, S. M., Shulyma, O. V., & Omelyanenko, K. A. (2013). Aktualnost modelirovanija raspredelennyh energosistem effektivnogo ispolzovanija vozobnovljaemyh istochnikov jenergii. Vostochno-Evropejskij zhurnal peredovyh tehnologij, 5(8), 4-8.

[17] Shulyma, O., Shendryk, V., Baranova, I., & Marchenko, A. (2014). The Features of the Smart MicroGrid as the Object of Information Modeling. Information and Software Technologies, 12. doi: 10.1007/978-3-319-11958-8_2

[18] The FIPA Abstract Architecture Specification. Retrieved March 16, 2015, from http://www.fipa.org/specs/fipa00001/index.html

Author Index