Position Papers of the 2016 Federated Conference on Computer Science and Information Systems

September 11–14, 2016. Gdańsk, Poland

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

PTI

Annals of Computer Science and Information Systems, Volume 9

Position Papers of the 2016 Federated Conference on Computer Science and Information Systems

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

2016, Warszawa,
Polskie Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 9

Position Papers of the 2016 Federated Conference on Computer Science and Information Systems

USB: ISBN 978-83-60810-94-1
WEB: ISBN 978-83-60810-93-4

ISSN 2300-5963
DOI 10.15439/978-83-60810-93-4

Ul. Solec 38/103
00-394 Warsaw
Poland

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

Cover:
Alisa Denisiuk,
Elbląg, Poland

Also in this series:

Volume 8: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-90-3, ISBN USB: 978-83-60810-91-0**
Volume 7: Proceedings of the LQMR Workshop, **ISBN WEB: 978-83-60810-78-1, ISBN USB: 978-83-60810-79-8**
Volume 6: Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-76-7, ISBN USB: 978-83-60810-77-4**
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 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Gdańsk, Poland, on September 11-14, 2016. This is the forth 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 2016 was Chaired by prof. Krzysztof Goczyła, while Zenon Filipiak acted as the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), Systems Research Institute Polish Academy of Sciences, Warsaw Univesity of Technology, Wrocław University of Economics, and Gdańsk University of Technology. It was organized in technical cooperation with: IEEE Region 8, IEEE SMC Technical Committee on Computational Collective Intelligence, IEEE Computer Society Technical Committee on Intelligent Informatics, IEEE Poland Section, Computer Society Chapter Poland, Gdańsk Computer Society Chapter Poland, Polish Chapter of the IEEE Computational Intelligence Society, ACM Special Interest Group on Applied Computing, Łódź ACM Chapter, European Alliance for Innovation (EAI), Committee of Computer Science of the Polish Academy of Sciences, Polish Operational and Systems Research Society, Mazovia Cluster ICT Poland and Eastern Cluster ICT Poland. Furthermore, the 11th International Symposium Advances in Artificial Intelligence and Applications (AAIA'16) 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.

FedCSIS 2016 consisted of the following events (conferences, symposia, workshops, special sessions). These events were grouped into FedCSIS conference areas**,** of various degree of integration. Specifically, those listed in italics and without indication of the year 2016 signify "abstract areas" with no direct paper submissions (i.e. paper submissions only within enclosed events).

- **AAIA'16 – 11th International Symposium Advances in Artificial Intelligence and Applications**
  - AIMaVIG'16 – 2nd International Workshop on Artificial Intelligence in Machine Vision and Graphics
  - AIMA'16 – 6th International Workshop on Artificial Intelligence in Medical Applications
  - AIRIM'16 – 1st International Workshop on AI aspects of Reasoning, Information, and Memory
  - ASIR'16 – 6th International Workshop on Advances in Semantic Information Retrieval
  - DSTUI'16 – 6th International Workshop on Dealing with Spatial and Temporal Uncertainty and Imprecision
  - LTA'16 – 1st International Workshop on Language Technologies and Applications
  - WCO'16 – 9th International Workshop on Computational Optimization
- ***CSS - Computer Science & Systems***
  - AIPC'16 – 1st International Workshop on Advances in Image Processing and Colorization
  - CANA'16 – 9th Computer Aspects of Numerical Algorithms
  - CPORA'16 – 1st Workshop on Constraint Programming and Operation Research Applications
  - IWCPS'16 – 3rd International Workshop on Cyber-Physical Systems
  - MMAP'16 – 9th International Symposium on Multimedia Applications and Processing
  - WSC'16 – 8th Workshop on Scalable Computing
- **ECRM – Education, Curricula & Research Methods**
  - IEES'16 – 1st International E-education Symposium - Education of the Future
  - DS-RAIT'16 – 3rd Doctoral Symposium on Recent Advances in Information Technology
- **iNetSApp'16 – 4th International Conference on Innovative Network Systems and Applications**
  - EAIS'16 – 3rd Workshop on Emerging Aspects in Information Security
  - SoFAST-WS'16 – 5th International Symposium on Frontiers in Network Applications, Network Systems and Web Services
  - WSN'16 – 5th International Conference on Wireless Sensor Networks
- ***IT4MBS – Information Technology for Management, Business & Society***
  - ABICT'16 – 7th International Workshop on Advances in Business ICT
  - AITM'16 – 14th Conference on Advanced Information Technologies for Management
  - ISM'16 – 11th Conference on Information Systems Management
  - KAM'16 – 22nd Conference on Knowledge Acquisition and Management
  - UHH'16 – 2nd International Workshop on Ubiquitous Home Healthcare
- ***JAWS – Joint Agent-oriented Workshops in Synergy***
  - MAS&S'16 – 10th International Workshop on Multi-Agent Systems and Simulations
  - SEN-MAS'16 – 4th International Workshop on Smart Energy Networks & Multi-Agent Systems
- **SSD&A – Software Systems Development & Applications**
  - BTMSPA'16 – 1st Symposium on Balancing Traditional and Modern Software Process Approaches
  - MDASD'16 – 4th Workshop on Model Driven Approaches in System Development

- MIDI'16 – 4th Conference on Multimedia, Interaction, Design and Innovation
- SEW-36 – The 36th IEEE Software Engineering Workshop

This year (2016) is the year of the 90th anniversary of the birth and the 10th anniversary of the death of Professor Zdzisław Pawlak. Therefore, a special plenary panel devoted to the "Legacy of Professor Zdzisław Pawlak" has been organized. During the panel, friends, students and collaborators of Prof. Pawlak shared their memories and reflections.

Furthermore, an AAIA'16 Data Mining Competition, focused on "Predicting Dangerous Seismic Events in Active Coal Mines" has been organized. Its results constitute a separate section in these proceedings. Awards for the winners of the contest were sponsored by: Research and Development Center EMAG and the Mazovia Chapter of the Polish Information Processing Society.

Each event constituting FedCSIS had its own Organizing and Program Committee. We would like to express our warmest gratitude to the members of all of them for their hard work attracting and later refereeing 512 submissions.

FedCSIS 2016 was organized under the auspices of dr. Jarosław Gowin, Minister of Science and Higher Education, Anna Streżyńska, Minister of Digital Affairs, Paweł Adamowicz, Mayor of the City of Gdańsk, Wojciech Szczurek Mayor of the City of Gdynia, and prof. Henryk Krawczyk, Rector of the Gdańsk University of Technology.

Finally, FedCSIS 2016 was sponsored by the Ministry of Science and Higher Eduction and Intel.

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*

Position Papers of the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS)

September 11–14, 2016. Gdańsk, Poland

TABLE OF CONTENTS

9TH International Workshop on Computational Optimization

Computer Science & Systems

1ST Workshop on Constraint Programming and Operation Research Applications

3RD International Workshop on Cyber-Physical Systems

8TH Workshop on Scalable Computing

Education, Curricula & Research Methods

11th International Symposium Advances in Artificial Intelligence and Applications

THE AAIA'16 will bring scientists, 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'16 will be of interest to researchers who want to know about both theoretical advances and latest applied developments in AI.

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:

- Decision Support
- Machine Learning
- Fuzzy Sets and Soft Computing
- Rough Sets and Approximate Reasoning (see also FedCSIS'16 Plenary Panel commemorating Prof. Zdzisław Pawlak)
- Data Mining and Knowledge Discovery
- Data Modeling and Feature Engineering
- Data Integration and Information Fusion
- Hybrid and Hierarchical Intelligent Systems
- Neural Networks and Deep Learning
- Bayesian Networks and Bayesian Reasoning
- Case-based Reasoning and Similarity
- Web Mining and Social Networks
- AI in Business Intelligence and Online Analytics
- AI in Robotics and Cyber-Physical Systems
- AI-centered Systems and Large-Scale Applications

We also encourage researchers interested in the following topics to submit papers directly to the corresponding workshops, which are integral parts of AAIA'16:

- AI in Computational Optimization (see WCO'16 workshop)
- AI in Language Technologies (see LTA'16 workshop)
- AI in Machine Vision and Graphics (see AIMaViG'16 workshop)
- AI in Medical Applications (see AIMA'16 workshop)
- AI in Reasoning and Computational Foundations (see AIRIM'16 workshop)
- AI in Semantic Information Retrieval (see ASIR'16 workshop)
- AI in Spatial and Temporal Analytics (see DSTUI'16 workshop)

All papers accepted to the main track of AAIA'16 and to the above workshops will be treated equally in the conference programme and will be equally considered for the awards listed below.

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.

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

We are proud to announce that, as in the recent years, the International Fuzzy Systems Association (IFSA) Best Paper Award for Young Scientist, will be presented.

Candidates for all the above awards can come from the AAIA'16 and all workshops organized within its framework.

EVENT CHAIRS

- **Janusz, Andrzej,** University of Warsaw, Poland
- **Ślęzak, Dominik,** University of Warsaw & Infobright Inc., Poland

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,** Wrocław University, Poland
- **Bazan, Jan,** University of Rzeszów, Poland
- **Bordogna, Gloria,** CNR IREA, Italy
- **Borkowski, Janusz,** Polish-Japanese Academy of Information Technology & Infobright Inc.
- **Błaszczyński, Jerzy,** Poznań University of Technology, Poland

- **Cetnarowicz, Krzysztof,** AGH University of Science and Technology, Poland
- **Chakraverty, Shampa,** Netaji Subhas Institute of Technology, India
- **Chen, Phoebe,** La Trobe University, Australia
- **Cheung, William,** Hong Kong Baptist University, Hong Kong S.A.R., China
- **Cyganek, Bogusław,** AGH University of Science and Technology, Poland
- **Czarnowski, Ireneusz,** Gdynia Maritime University, Poland
- **Czerniak, Jacek M.,** Casimir the Great University in Bydgoszcz, Poland
- **Czyżewski, Andrzej,** Gdańsk University of Technology, Poland
- **Dardzińska, Agnieszka,** Białystok University of Technology, Poland
- **Dey, Lipika,** Tata Consulting Services, India
- **do Carmo Nicoletti, Maria,** UFSCar & FACCAMP, Brazil
- **Duentsch, Ivo,** Brock University, Canada
- **Eklund, Patrik,** Umeå University, Sweden
- **Froelich, Wojciech,** University of Silesia, Poland
- **Holzinger, Andreas,** Graz University of Technology, Austria
- **Jatowt, Adam,** Kyoto University, Japan
- **Jin, Xiaolong,** Institute of Computing Technology of Chinese Academy of Sciences, China
- **Kayakutlu, Gulgun,** Istanbul Technical University, Turkey
- **Korbicz, Józef,** University of Zielona Góra, Poland
- **Kostek, Bożena,** Gdańsk University of Technology, Poland
- **Krasuski, Adam,** Main School of Fire Service (SGSP), Poland
- **Kryszkiewicz, Marzena,** Warsaw University of Technology, Poland
- **Lopes, Lucelene,** PUCRS, Brazil
- **Madalińska-Bugaj, Ewa,** University of Warsaw
- **Marek, Victor,** University of Kentucky, United States
- **Matson, Eric T.,** Purdue University, United States
- **Menasalvas, Ernestina,** Universidad Politécnica de Madrid, Spain
- **Mercier-Laurent, Eunika,** University Jean Moulin Lyon3, France
- **Mirkin, Boris,** Birkbeck University of London & NRU Higher School of Economics in Moscow, Russia
- **Miyamoto, Sadaaki,** University of Tsukuba, Japan
- **Moshkov, Mikhail,** King Abdullah University of Science and Technology, Saudi Arabia
- **Musiał-Gabryś, Katarzyna,** Bournemouth University, United Kingdom
- **Myszkowski, Paweł,** Wrocław University of Technology, Poland
- **Nguyen, Hung Son,** University of Warsaw, Poland
- **Nourani, Cyrus F.,** Akdmkrd-DAI TU Berlin & Munich Transmedia & SFU Burnaby, Germany
- **Nowostawski, Mariusz,** Norwegian University of Technology and Science (NTNU), Norway
- **Ohsawa, Yukio,** University of Tokyo, Japan
- **Peters, Georg,** Munich University of Applied Sciences, Germany
- **Po, Laura,** Università di Modena e Reggio Emilia
- **Porta, Marco,** University of Pavia, Italy
- **Proficz, Jerzy,** Academic Computer Center, Gdansk University of Technology, Poland
- **Przybyła-Kasperek, Małgorzata,** University of Silesia, Poland
- **Raghavan, Vijay,** University of Louisiana at Lafayette, United States
- **Rao, Raghavendra,** University of Hyderabad, India
- **Rauch, Jan,** University of Economics, Prague, Czech Republic
- **Reformat, Marek,** University of Alberta, Canada
- **Ruta, Dymitr,** Khalifa University, United Arab Emirates
- **Ryżko, Dominik,** Warsaw University of Technology, Poland
- **Santofimia, Maria Jose,** Universidad de Castilla-La Mancha, Spain
- **Schaefer, Gerald,** Loughborough University, United Kingdom
- **Sikora, Marek,** Silesian University of Technology, Poland
- **Su, Chang,** Chongqing University of Posts and Telecommuncations
- **Sydow, Marcin,** Polish Academy of Sciences & Polish-Japanese Academy of Information Technology, Poland
- **Szczęch, Izabela,** Poznań University of Technology, Poland
- **Szczuka, Marcin,** University of Warsaw, Poland
- **Szpakowicz, Stan,** University of Ottawa, Canada
- **Szwed, Piotr,** AGH University of Science and Technology, Poland
- **Tsay, Li-Shiang,** North Carolina A&T State University, United States
- **Unland, Rainer,** Universität Duisburg-Essen, Germany
- **Unold, Olgierd,** Wrocław University of Technology, Poland
- **Wang, Xin,** University of Calgary, Canada
- **Weber, Richard,** Universidad de Chile, Chile
- **Wieczorkowska, Alicja,** Polish-Japanese Academy of Information Technology, Poland
- **Woźniak, Michał,** Wrocław University of Technology, Poland
- **Wróblewski, Jakub,** Infobright Inc.
- **Zadrożny, Sławomir,** Systems Research Institute of Polish Academy of Sciences, Poland
- **Zakrzewska, Danuta,** Łódź University of Technology, Poland
- **Zielosko, Beata,** University of Silesia, Poland
- **Ziółko, Bartosz,** AGH University of Science and Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 3–9
DOI: 10.15439/2016F371
ACSIS, Vol. 9. ISSN 2300-5963

Clustering Validity Indices evaluation with regards to semantic homogeneity

Tomasz Dziopa
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Tomasz.Dziopa@students.mimuw.edu.pl

***Abstract*—Clustering validity indices are methods for examining and assessing the quality of data clustering results. Various studies provide a thorough evaluation of their performance using both synthetic and real-world datasets. In this work, we describe various approaches to the topic of evaluation of a clustering scheme. Moreover, a new solution to a problem of selecting an appropriate clustering validity index is presented. The approach is applied to a problem of selecting a suitable clustering validity index for a real-world task of clustering biomedical articles using the MeSH ontology.**

I. Introduction

THE PROBLEM of clustering is one of the fundamental problems in Machine Learning. The goal of clustering is to find the best way to divide a set of points into groups. This formulation reflects a natural process of learning—humans tend to categorize entities, like objects, people or events into clusters, which are characterized by common attributes. In this paper, we will present a comparison of clustering validity indices with regards to their applications to clustering biomedical documents from PubMed database.

Clustering validity is a common name for quantitative evaluation of the results of clustering algorithms [12]. Clustering Validity Index (CVI) can be perceived as a function which takes as arguments the dataset and clustering scheme and outputs some value which represents the quality of the clustering scheme.

Cluster validity index should provide some insight about the quality of grouping. The most intuitive notions reflected by the concept of "good clustering" are *compactness* and *separation*. The cluster is compact, when points within this cluster are possibly close to each other, whereas clusters are separated, when neighboring clusters are possibly far from each other [4].

The most common applications of cluster validity methods are:

1) Fine-tuning parameters of clustering parameters—the comparison of varying clustering schemes obtained using different parameters in order to find the best grouping.
   One of the most common parameters studied in the literature is the number of clusters in algorithms that assume fixed number of clusters *a priori*, like k-means citeHennig2014.
2) Examining *clustering stability* of a dataset—sensitivity of result of clustering algorithm to modification of algorithm's parameters
3) Examining *clustering tendency*—in some cases we do not know if the dataset has any clustering structure so that it can be grouped in a meaningful way. By applying cluster validity methods we can determine, whether the dataset has adequate grouping structure.

II. Clustering Validity Indices

Clustering Validity Indices are most commonly categorized into three main categories: internal, external and relative. In this chapter we will present the most common methods for assessing the quality of a clustering scheme.

A. External methods

Indices from this group assume that for dataset D some reference clustering $T = \{T_1, \ldots, T_m\}$ is given. The idea is that these indices try to express similarity between some scheme $C = \{C_1, \ldots, C_k\}$ being examined and T, sometimes referred to as *gold standard*.

Pair-counting indices

We introduce a label for a pair of points (x_a, x_b) for each $x_a, x_b \in D$:

- True Positives: x_a and x_b belong to the same partition in T and are also in the same partition in C.
- False Negatives: x_a and x_b belong to the same partition in T, but are in different partitions in C.
- False Positives: x_a and x_b do not belong to the same partition in T, but they belong to the same partition in C.
- True Negatives: x_a and x_b belong to different partitions in both T and C.

Pair-counting indices are defined as functions calculated over the sizes of the TP, TN, FP and FN sets.

1) Rand Statistic:

$$R = \frac{TP + TN}{TP + TN + FP + FN}$$

Index describes the ratio of correctly guessed pairs (clusterings C and T agree on membership of both points to either the same or different clusters). Perfect clustering will achieve $R = 1$.

2) Jaccard Coefficient:

$$J = \frac{TP}{TP + FN + FP}$$

Perfect clustering achieves $J = 1$, as there are no false negatives and no false positives. Jaccard coefficient is asymmetric in terms of true negatives and true positives, as it ignores true negatives. The influence of pairs of points belonging to the same cluster in both clusterings is amplified and the impact of pairs of points not belonging together is discounted.

3) Fowlkes and Mallows index: Let's introduce the notions of *pairwise precision* and *pairwise recall*, defined as follows:

$$prec = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

The Fowlkes-Mallows index is defined as the geometric mean of the *pairwise precision* and *pairwise recall*:

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

This measure is asymmetric in terms of true positives and true negatives, because true negatives are ignored. Maximum value of FM is 1, when there are no false positives or negatives.

Matching-based measures

Matching-based measures try to match clusters from C with *gold standard* clusters T and calculate various statistics on the matching.

4) Purity: This measure tries to capture the concept of cluster being *pure* - that is, containing only points from one golden-standard partition. Purity can be defined as follows[13]:

$$purity = \frac{1}{N} \sum_i \max_j |C_i \cap T_j|$$

where T is the set of ground-truth clusters and C is the set of examined clusters.

We can distinguish following cases based on the cardinality of sets C and T:

1) when $|C| = |T|$ and purity $= 1$, then C is a perfect clustering
2) when $|C| > |T|$ purity can still achieve 1, when each cluster in C is a subset of cluster in ground-truth partitioning T
3) when $|C| < |T|$ purity is always < 1 - at least one cluster in C contains points from more than one clusters in T

5) Maximum matching: Maximum matching [13] is defined as the value of maximum matching between sets in C and T - unlike in purity, each cluster in C is assigned a unique partition from T.

More formally, given a graph $G = (V, E)$, where $V = C \cup T$ and $\forall_{i,j}(C_i, T_j) \in E$ we want to find a maximum weighted matching in G. Weights on edges are given as $w(C_i, T_j) = |C_i \cap T_j|$.

The problem of finding a maximum matching in a bipartite weighted graph, assuming $|C| \approx |T|$, can be solved in $O(|C|^2 \log |C| + |C|^3) = O(|C|^3)$ time complexity.

6) F-Measure: Let $T_{match}(i) = \arg\max_{T_j \in T} |C_i \cap T_j|$ denote the cluster in ground-truth partition T, which is represented the most in cluster C_i. We can define precision and recall for cluster C_i as follows:

$$prec_i = \frac{|C_i \cap T_{match}(i)|}{|C_i|}$$

$$recall_i = \frac{|C_i \cap T_{match}(i)|}{|T_{match}(i)|}$$

The F-measure for cluster C_i is a harmonic mean of the precision and recall for this cluster:

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i}$$

F-measure for the clustering scheme C is the mean of F-measures for all clusters:

$$F = \frac{1}{|C|} \sum_{i=1}^{|C|} F_i$$

B. Internal methods

Internal indices are designed to express some properties of the resulting clustering scheme with regards to proximity measure.

Internal methods operate on the proximity matrix, which can be defined as:

$$W = \{\delta\{x_i, x_j\}\}_{i,j=1}^{n}$$

Proximity measure δ should be non-negative, symmetrical and fulfill the triangle inequality.

1) Dunn index: Dunn index is defined as a ratio of the minimum distance between clusters to the maximum cluster's diameter. These two notions can be interpreted in various ways, resulting in various definitions of Dunn Index.

Inter-cluster distance can be defined as:

- minimum distance between points originating from different clusters,
- maximum distance between points originating from different clusters,
- distance between centroids of the clusters

Cluster's diameter can be defined as:

- maximum distance between two points within the cluster,
- mean distance between all pairs of points from the cluster,
- sum of distances of each points to the mean of the cluster

The larger the Dunn index, the better the clustering - the distance between points in different clusters is much larger than the distance between points inside the same cluster. However, Dunn index can be insensitive as inter- and intracluster distance does not capture all information about the clustering.

2) Davies-Bouldin Index: Let μ_i denote the mean of cluster C_i:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

and σ_i denote the dispersion of the points in the cluster C_i around it's mean μ_i:

$$\sigma_i = \sqrt{\frac{\sum_{x_j \in C_i} \delta(x_j, \mu_i)}{|C_i|}}$$

The Davies-Bouldin measure [6] for pair of clusters C_i, C_j is defined as follows:

$$DB_{ij} = \frac{\sigma_i + \sigma_j}{\delta(\mu_i, \mu_j)}$$

DB_{ij} measures the compactness of clusters compared to the distance between the cluster means.

$$DB = \frac{1}{|C|} \sum_{i=1}^{|C|} \max_{j \neq i} \{DB_{ij}\}$$

That is, for each cluster C_i we pick another cluster C_j which produces the largest value of DB_{ij} ratio. The smaller the DB value the better the clustering, because this means that clusters are well-separated (the distance between cluster means is large) and each cluster is compact (has a small spread).

3) Silhouette Coefficient: Silhouette coefficient [11] is a measure of both compactness and separation of clustering. Let a_i denote average dissimilarity of x_i with all other points within its cluster. a_i can be interpreted as how well x_i has been assigned to its cluster. Let b_i denote the lowest average dissimilarity of x_i to any other cluster in C, of which x_i is not a member. Assuming $x_i \in C_j$:

$$a_i = \frac{1}{|C_j|} \sum_{x_l \in C_j; x_l \neq x_i} \delta(x_i, x_l)$$

$$b_i = \min_{C_l \in C; C_l \neq C_j} \frac{1}{|C_l|} \sum_{x_k \in C_l} \delta(x_i, x_k)$$

The silhouette coefficient for data point x_i is defined as:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

s_i can obtain values in interval $[-1, 1]$. $s_i = 1$ indicates that x_i is close to points in its assigned cluster and far from other clusters, $s_i = 0$ indicates that x_i lies close to the boundary between two neighbouring clusters. $s_i = -1$ indicates that x_i is much closer to another cluster than its own cluster - the point has been misclustered.

Silhouette coefficient for clustering C is defined as:

$$SC = \frac{1}{N} \sum_{i=1}^{N} s_i$$

4) Normalized Γ: Let W be the proximity matrix of the dataset, and Y be the proximity matrix defined as follows:

$$Y = \left\{\delta(\mu_{x_i}, \mu_{x_j})\right\}_{i,j=1}^{n}$$

μ_{x_i} is the mean of all points that belong to the same cluster as x_i. Let $\mathbf{w}, \mathbf{y} \in \mathbb{R}$ be vectors obtained by linearizing the upper triangular elements excluding main diagonal of W and Y.

Let $\mathbf{z}_W$ and $\mathbf{z}_Y$ denote mean-centered vectors $\mathbf{w}, \mathbf{y}$. Now, the normalized Γ statistic can be defined as:

$$\Gamma_n = \frac{\mathbf{z}_W^T \mathbf{z}_Y}{||\mathbf{z}_W|| \cdot ||\mathbf{z}_Y||}$$

5) Within-Between Ratio: Within-Between Ratio is a ratio of average distance within clusters μ_{within} to average distance between clusters $\mu_{between}$.

$$\mu_{within} = \frac{\sum_{C_i \in C} \sum_{x_j, x_k \in C_i; j \neq k} \delta(x_j, x_k)}{\sum_{C_i \in C} \binom{|C_i|}{2}}$$

$$\mu_{between} = \frac{\sum_{C_i, C_j \in C; i \neq j} \delta(C_i, C_j)}{\binom{|C|}{2}}$$

$$WB = \frac{\mu_{within}}{\mu_{between}}$$

The smaller Within-Between Ratio, the better the clustering scheme.

6) Within cluster sum of squares: Within cluster sum of squares is a sum of within-cluster squared dissimilarities divided by the cluster size.

$$WCSS = \frac{1}{2} \sum_{C_i \in C} \frac{\sum_{x_j, x_k \in C; j \neq k} \delta(x_j, x_k)}{|C_i|}$$

The smaller the Within Cluster Sum of Squares, the more compact are the clusters.

7) Calinski-Harabasz Index: Given a clustering of a dataset $C = \{C_1, \ldots, C_k\}$ consisting of N points, Calinski-Harabasz index is defined as [3]:

$$CH(k) = \frac{SS_B}{SS_W} \cdot \frac{N - k}{k - 1}$$

SS_B is the overall between-cluster variance, defined as:

$$SS_B = \sum_{i=1}^{k} |C_i| \cdot ||\mu_i - \mu||^2$$

where μ_i is a mean of i-th cluster, μ is an overall mean of the sample data, and $||\mu_i - \mu||$ is the L^2 norm (Euclidean distance) between the two vectors.

SS_W is the overall within-cluster variance, defined as:

$$SS_W = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

The larger the value of Calinski-Harabasz index, the better the quality of the clustering scheme - good clustering has large between-cluster variance SS_B and a small within-cluster variance SS_W.

C. Relative Validity Indices

Relative validity indices are used for comparison of clustering schemes. Indices from this group are used for deciding which clustering scheme fits the data best. This definition covers both external and internal indices which can assess the quality in relative terms, however, this notion usually refers to internal methods that are also relative.

The most common use case for relative validity indices is selecting the best clustering scheme from the set of schemes obtained using different parameters.

III. Other approaches to Clustering Validity

A. Ensembles of Clustering Validity Indices

Internal indices presented in the previous section tend to capture only particular properties of evaluated clustering scheme. The question that arises is whether a combination of internal indices would correctly judge the quality of every clustering scheme.

In [9] multiple strategies of building ensembles of clustering validity indices are evaluated. Authors have conducted an experiment on various synthetic and real-life datasets and examined the correlation of the ensembles of relative indices with regards to external validity index. In this work the quality of validity index (or ensemble) is measured as Spearman's rank correlation coefficient with external index, referred to as *effectiveness*.

The basic motivation for validation using ensembles of CVI is that multiple indices with high effectiveness and a high degree of complementarity should produce more robust results than any single index.

Another problem regarding ensembles of CVI is the choice of aggregation strategy. In [9] authors have compared various strategies of combining the results of the indices. The authors found, that score-based strategies, like mean or median of the normalized values of indices in the ensemble, appeared to give inferior results to methods based on rank aggregation. The main advantage of using rank-based aggregation is that it does not rely on the concrete values of indices, therefore it does not require value normalization.

The conclusions from the experiments are, that all examined ensembles, even those assembled from random subsets of measures and with random aggregation strategy, achieved higher effectiveness than the expected value of single validity index. Surprisingly, the strategy of aggregation of results did not affect the overall effectiveness of the ensembles.

B. Semantic Approach to Clustering Validity

We can divide the indices describe previously into two major groups: internal and external. Internal indices rely only on problem space but capture the specific characteristic of the clustering scheme. External indices rely only on explicit clustering assignment therefore limiting the expedience in real world scenarios. Another approach to clustering scheme validation is the use of partial information, which does not reflect direct cluster membership, but rather implies the semantic relationships between the points.

More formally, apart from the dataset D we are given a set $L = \{L_1, \ldots, L_n\}$ consisting of sets of labels with a one-to-one correspondence between elements from L and D.

In this approach, the clustering is produced based only on the original space, discarding the additional information. This information is used afterward to assess the quality of the output clustering.

The formulation of the clustering validation problem is motivated by the rise of popularity of datasets which are annotated using labels, *tags*, or *hashtags*.

Semantic Explorative Evaluation

Semantic Explorative Evaluation described in [10] tries to capture human reasoning of assessing clustering scheme. When an expert faces the problem of manual evaluation of clustering results, he tries to explain the contents of the clusters in his own words. The main idea of SEE is that the quality of the cluster is correlated with the measure of the complexity of expert description of the cluster.

More formally, SEE takes as an input the dataset tagged with expert tags describing the points in the dataset and the clustering. Next, for each cluster, we need to calculate the complexity of expert's description of the cluster - a model of the cluster in terms of expert tags. Any classification algorithm can provide such a model. Then the model complexity measure needs to be defined. Authors have chosen decision tree classifier as their classification algorithm, and an average depth of the resulting tree as the measure of model's complexity.

IV. Evaluation of Cluster Validity Indices

In most works evaluating CVIs, the first step is to choose a set of datasets. Usually, synthetic datasets are used forcing various characteristics of desired clustering schemes, like varying densities, compactness, overlapping, shapes, or added noise. Additionally, a number of sample real-world datasets with known number of clusters can be chosen for the experiments.

A. Optimal K criterion

Cluster validity index evaluation has been thoroughly covered in many papers in recent years. Authors of [5] evaluated multiple papers on the topic. Surprisingly, multiple works appear to use the same methodology.

The methodology requires preparation of synthetic datasets with known number of clusters. For this reason, usually, two-dimensional datasets are used, for the ease of visualization and human verification. Additionally, we need to choose a clustering algorithm for the experiment, which allows an input parameter that sets the number of clusters for the output partition, k. The most popular algorithms in the literature are agglomerative hierarchical algorithm and k-means.

Let's denote the ideal partition of a dataset as P^*. Subsequently, algorithm is run over the dataset with a set $K = \{k_1, \ldots, k_l\}$ of different values of parameter k. As a result, we obtain a set of partitions, $S = \{P_1, \ldots, P_l\}$, with one of them being a partition with a correct number of clusters for the dataset, denoted as P_N. More formally,

$$P_N = \{P_i \in P : |P^*| = |P_i|\}$$

Finally, CVI is computed for all partitions in S. The idea is, that the partition obtaining the best value for the evaluated $CVI(P_x)$ will serve to predict an actual number of clusters. Let's assume for simplicity, that function $CVI(P_x)$ assigns greater values to "better" partitions. We say the partition P_{CVI} is proposed by the cluster validity index, when

$$P_{CVI} = \arg\max_{P_i \in S} CVI(P_i)$$

Clustering validity index has predicted that the dataset contains $|P_{CVI}|$ clusters if it has made a successful guess so that $|P_{CVI}| = |P^*|$.

The method works under a fundamental assumption, that algorithm used for clustering works "correctly" - that is, algorithm-generated partition P_N is the one that fits the data best. Obtained results of CVI are biased if the assumption does not hold so that there exists partition P_i that captures the clustering scheme of the data better.

B. External criteria similarity

The problem of unrealistic assumption of the clustering algorithm being able to correctly partition every dataset has been addressed in [5]. The authors have proposed a modified version of the Optimal k criterion. In contrast to this method, the CVI is said to have succeeded if it has proposed the partition most similar to the optimal partitioning, instead of the partition containing the same number of clusters as an ideal partition.

Similar as in previous method, we need to provide the input dataset D, the set of potential values of parameter k, $K = \{k_1, .., k_l\}$, the set of partitions from a clustering algorithm $S = \{P_1, \ldots, P_l\}$ and the *gold standard* partition P^*. Additionally, we need to provide a partition similarity measure $sim(P_i, P_j)$, for example one of external validity indices. Then, the partition obtaining highest similarity of all computed partitions can be defined as:

$$\widehat{P} = \arg\max_{P_i \in S} sim(P_i, P^*)$$

In the new methodology, we say clustering validity index has made a successful guess if $P_{CVI} = \widehat{P}$ - when the partition obtaining the best value of examined CVI is at the same time the most similar to the ideal partitioning.

The similarity measure is another input parameter of the methodology, so its choice can be adapted to the characteristics of the experiment. Extension of this idea is to use multiple partition similarity measures and to either aggregate their results by averaging or using a voting system.

V. Semantic Evaluation of Clustering Validity Indices

We present a new approach to selecting the best clustering validity index. We examine the problem of clustering with additional information. The intuition behind this problem formulation is that we are given a data set, which has been manually annotated with multiple labels reflecting semantic relationships between documents. The clustering is produced based only on the original space, discarding the additional information. This information is used afterward to assess the quality of the output clustering.

More formally, apart from the data set D we are given a set $L = \{L_1, \ldots, L_n\}$ consisting of sets of labels with a one-to-one correspondence between elements from L and D.

The motivation behind this formulation of the clustering validation problem is, that it uses additional information about the relationships, which is not an explicit grouping of the points. Moreover, recently the number of data sets annotated using labels, *tags*, or their social-network equivalents, *hashtags* has increased.

The proposed method requires calculating the partitionings S of the dataset using only information from D into k groups, for each k in $K = \{k_1, \ldots, k_l\}$. Similarly as in Section IV-B, we want to assess the quality of the clustering using external knowledge, but since we do not have reference clustering, we cannot use an external CVI. Instead, we calculate the semantic quality index $ASH(P_i)$, proposed in [8].

The ASH index uses a notion of *semantic distance*, which is defined for documents T_i, T_j with corresponding sets of assigned expert tags L_i, L_j to be a F1 score between sets L_i, L_j:

$$F_1 distance(T_i, T_j) = 1 - 2 \cdot \frac{precision(L_i, L_j) \cdot recall(L_i, L_j)}{precision(L_i, L_j) + recall(L_i, L_j)} \quad (1)$$

Precision and recall are defined as:

$$precision(L_i, L_j) = \frac{|L_i \cup L_j|}{|L_i|} \quad (2)$$

$$recall(L_i, L_j) = \frac{|L_i \cap L_j|}{|L_j|} \quad (3)$$

The *semantic distance* between two sets of documents is defined as an average of pairwise $F_1 distances$ between pairs of texts from different sets:

$$semDist(D_1, D_2) = \frac{\sum_{T_i \in D_1, T_j \in D_j} F_1 distance(T_i, T_j)}{|D_1| \cdot |D_2|} \quad (4)$$

The measure of document's semantic homogeneity is defined similarly as Silhouette Coefficient:

$$homogeneity(T_i) = \frac{B(T_i) - A(T_i)}{\max(A(T_i), B(T_i))} \quad (5)$$

$A(T_i)$ is a $semDist$ distance calculated between document T_i and all other documents within the same cluster as T_i. $A(T_i)$ can be interpreted as the measure of quality of T_i's assignment to its cluster. In case T_i forms a singleton cluster, $A(T_i) = 0$.

$B(T_i)$ is the $semDist$ measure calculated between document T_i and all documents which do not belong to the same cluster as T_i. $B(T_i)$ is the dissimilarity measure describing how far a document T_i is from all other clusters.

Finally, we define the Average Semantic Homogeneity:

$$ASH = \frac{1}{|D|} \sum_{T_i \in D} homogeneity(T_i) \quad (6)$$

as the measure of semantic quality of clustering scheme.

Similarly as in a methodology described in Section IV-B, we could formulate the CVI's correctness criterion as:

$$\arg\max_{P_i \in S} ASH(P_i) = P_{CVI} \quad (7)$$

Stating that the partition obtaining the highest value of Average Semantic Homogeneity is the one which is suggested by examined CVI.

However, in real world scenarios the size and dimensionality of the datasets may be too big for this criterion to select one particular clustering as the one fitting the data best.

We propose a modified CVI correctness criterion: we say that CVI is suitable for the task of evaluation the clustering schemes and preserves semantic relationships between the documents, when:

- CVI calculated on document space indicates optimal number of clusters
- CVI calculated on document space is correlated with Average Semantic Homogeneity index

The main advantage of this method is that it can be used for real-world applications since it does not require a ground-truth partitioning for the dataset. Instead, we pick a random sample from the dataset and have it manually annotated by the experts. Then, using the described method we select the good CVI for assessing the quality of the clustering of the subset of the original dataset. Finally, we state that the selected CVI is appropriate for assessing the quality of the original dataset.

VI. The Experiment

We have conducted an experiment to demonstrate the usage of Semantic Evaluation of Clustering Validity Indices in order to find the best CVI for assessing the quality of clustering text documents.

In the experiment, we used a dataset obtained from U.S. National Library of Medicine (NLM). The dataset consists of 42200 abstracts of scientific articles in English. Each article has been manually labeled by experts from NLM using concepts from MeSH ontology [1] using on average 12 concepts.

Abstracts were tokenized, stemmed and common English stopwords were removed. Documents are modeled using bag-of-words in a document-term matrix with tf-idf weighting [2]. MeSH terms are treated as labels, discarding the information about major topics and contexts in which the term appears, resulting in the total of 17169 unique labels.

The experiment has been conducted on a subset of 5054 documents from NLM dataset which were selected from the MeSH headings presented in the Table I. The selection of thematically disjoint headings is intended to strengthen the clustering tendency of the dataset.

The values presented on Figure 1 are mean aggregate on 10 randomly chosen without replacement, equinumerous subsets ($|D| \approx 505$). The datasets have been partitioned with Agglomerative Nesting algorithm implementation using cosine distance. The values of indices were calculated using `cluster.stats` implementation from R package `fpc` [7].

Figure 2 shows the values of Average Semantic Homogeneity - the value of semantic quality of clustering. Average Semantic Homogeneity achieves higher values for smaller clusters, with a maximum value of 1 for singleton clusters. We can find the best value of parameter k using the elbow criterion, with possible values of $k = \{5, 9, 14, 22\}$.

TABLE I
Characteristics of dataset used for experiment

MeSH heading	No. of documents
Bacterial proteins	772
Brain	652
Breast Neoplasms	848
Pregnancy	1033
E. Coli	574
HIV	629
Malaria	251
Diabetes	491

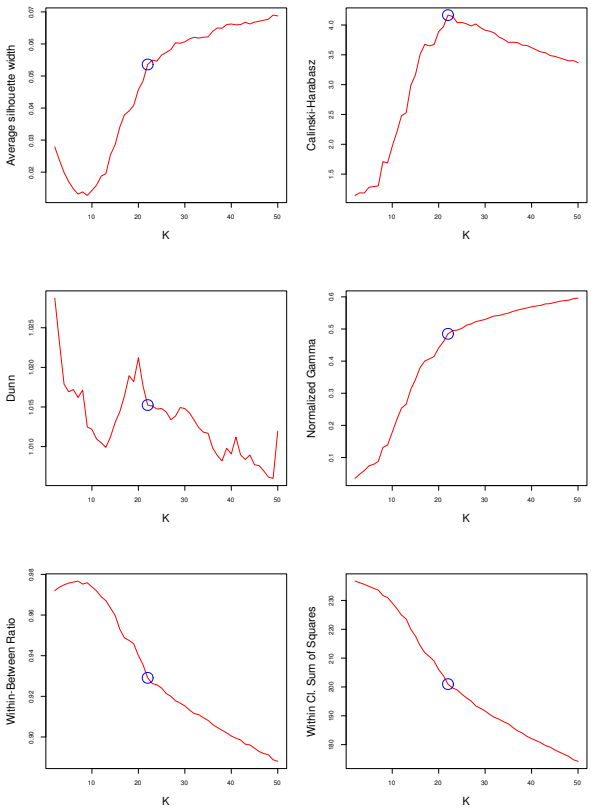


Fig. 1. Values of CVI indexes calculated by Average silhouette width, Calinski-Harabasz, Dunn, Normalized Γ, Within-Between Ratio and Within Cluster Sum of Squares methods on clusterings into $k \in [2, 50]$ groups in document space. Values for optimal number of clusters $k = 22$ have been marked with blue circle.

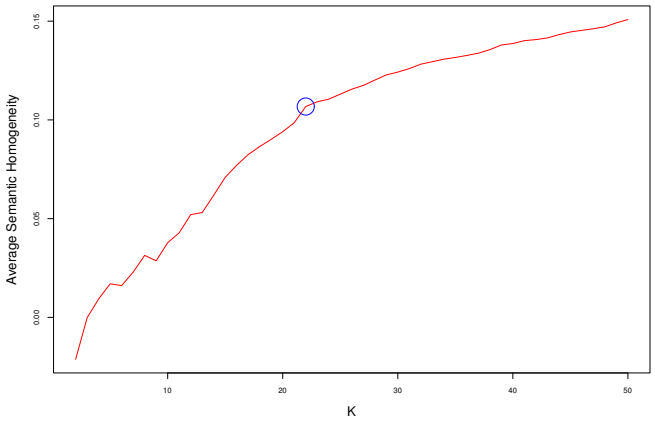


Fig. 2. Value of Average Semantic Homogeneity in experiment setup with optimal number of clusters $k = 22$ marked with blue circle

TABLE II
THE RESULTS OF COMPARISON

Index	L_2 Norm	best k
Calinski-Harabasz	1.200	22
Average Silhouette Width	1.014	22 (elb. crit.)
Dunn	2.102	20 (local max.)
Normalized Γ	0.184	18, 22 (elb. criterion)
Within-Between Ratio	1.889	-
Within Cl. Sum of Squares	1.063	-

The results of the experiment are summarized in Table II. The L_2 norm has been calculated on 0-1 normalized values of Average Semantic Homogeneity and examined clustering validity indices. Our experiment shows, that Normalized Γ shows very high correlation with ASH. Moreover, the index enables to find the best value of parameter k for evaluated dataset.

The Average Silhouette Width and Calinski-Harabasz index show relatively high correlation with ASH and both reach the optimal value of $k = 22$. Moreover, Calinski-Harabasz index achieves global maximum at $k = 22$, additionally strengthening the argument of 22 being the optimal value of k.

Dunn index has a relatively weak correlation with ASH, although it has the local optimum at $k = 20$. We might suppose, that it does not preserve semantic relationships between the documents, and suggested value is derived from other properties of the dataset.

The remaining indices, Within-Between Ratio and Within Cluster Sum of Squares do not suggest an optimal number of partitions for the dataset.

VII. CONCLUSIONS

In this work, we have shown that for the given problem, Calinski-Harabasz, Average Silhouette Width and Normalized Γ indices appear to reflect semantic relationships between the clusterings using bag-of-words and labels annotations models.

In contrast to previous studies, the method does not make any assumption on the correctness of the clustering algorithm. Moreover, this approach does not require datasets with known ground-truth partitioning. The presented methodology using the semantic measure of clustering scheme can be used in real-life problems. Additionally, the number of datasets tagged with expert labels or ontologies has increased in recent years.

One of the many possible directions for development of this method is to evaluate other measures of the semantic quality of the clustering. Additionally, we could make more extensive use of the MeSH ontology. In this thesis, we treated the concepts as labels, but we can take advantage of the tree-like structure hierarchy of MeSH terms and incorporate this knowledge into distance calculation in the label representation of data.

Furthermore, future studies should investigate the applications of the method to other types of documents. In this work we have examined the usage of the method with scientific documents from a particular domain only, whereas the overall applicability to clustering other kinds of documents should be researched. It is also worth examining the applications of the ensembles [9] of multiple CVIs pointed out as suitable for assessing the document clusters by our method.

REFERENCES

[1] https://www.nlm.nih.gov/mesh/introduction.html, 2016. [Online; accessed 5.05.2016].
[2] Charu C. Aggarwal and Cheng Xiang Zhai. *Mining Text Data*. Springer Publishing Company, Incorporated, 2012.
[3] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27, 1974.
[4] Brian Everitt. Cluster analysis. *Quality and Quantity*, 14(1):75–100, 1980.
[5] Ibai Gurrutxaga, Javier Muguerza, Olatz Arbelaitz, Jesús M. Pérez, and José Ignacio Martín. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*, 32(3):505–515, 2011.
[6] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.
[7] Christian Hennig. *fpc: Flexible Procedures for Clustering*, 2015. R package version 2.1-10.
[8] Andrzej Janusz, Dominik Ślęzak, and Hung Son Nguyen. Unsupervised similarity learning from textual data. *Fundam. Inf.*, 119(3-4):319–336, August 2012.
[9] Pablo A. Jaskowiak, Davoud Moulavi, Antonio C. S. Furtado, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems*, pages 1–26, 2015.
[10] Hung Son Nguyen, Sinh Hoa Nguyen, and W. Swieboda. Semantic explorative evaluation of document clustering algorithms. In *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, pages 115–122, Sept 2013.
[11] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.
[12] Sergios Theodoridis and Konstantinos Koutroumbas. Chapter 16 - cluster validity. In Sergios Theodoridis, , and Konstantinos Koutroumbas, editors, *Pattern Recognition (Fourth Edition)*, pages 863 – 913. Academic Press, Boston, fourth edition edition, 2009.
[13] Mohammed J. Zaki and Jr. Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, May 2014.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 11–17
DOI: 10.15439/2016F73
ACSIS, Vol. 9. ISSN 2300-5963

Many-valued logic in manufacturing

Patrik Eklund
Umeå University
Department of Computing Science
SE-90187 Umeå, Sweden
Email: peklund@cs.umu.se

Magnus Löfstrand
Umeå University
Department of Computing Science
SE-90187 Umeå, Sweden
Email: maglof@cs.umu.se

***Abstract*—This paper shows how to enrich the language used in the manufacturing industry regarding information structure and its representation for products and production processes. This is enabled because of our use of many-valued logic, and in order to complement the numerical approach commonly appearing in such representations. We underline the importance of utilizing mathematical disciplines like algebra and logic side-by-side with the utility of analysis and stochastics.**

I. Introduction

WITHIN a communicating resources perspective, many-valued logic will be applied to decrease Data *Complexity* for applications in *Big* Data. We will further introduce a structure for *functioning* classification in order to complement and interact with the traditional view of *faults and failures* in product and production subsystems. We will develop a prototype information structure demonstrating the potential use of a multi-valued logic enriched classification of *functioning in machines and manufacturing* (MCFu), as related to enriched classification of *faults in machines and manufacturing* (MCFa). Our *concept of digitalization*, in order to support manufacturing and machine simulation and modelling as e.g. part of design processes, whether e.g. in the case of cars and car production, related to subsystems such as powertrain, electronics or interior, will be based on nomenclatures, classification and ontology as part of production and in particular production where *information and process* is tightly connected. Our language and concept, with MCFu and MCFa classifications, will support transformation and transferal of data between robots, humans and companies (Internet of Things 2.0), thus strengthening industry.

For machines and vehicles, **faults** and **failures** reduce their function either partly or completely. The traditional enngineering view of functioning classifications does not always connect well and accurately with corresponding faults classifications. Machines and vehicles in the widest sense, e.g., including operators and end-users of a wide variety, need to interact with the environment, and therefore also need to appropriately connect *all resources*. To facilitate the connections, these respective resources need a *common language for representation of faults and functioning* and the relation between them. Further, within and across various machine subsystems, information and information structures are currently insufficiently connected with respect to standard information representation. Numerical approaches to standard and information structure development and utility within the industry promotes uncertainty and *many-valued considerations* mostly in directions of analyzing variability, and indeed variability as related to numerical values. Logical many-valuedness, and underlying structures are basically missing in most numerical approaches.

Our approach to many-valued logic can be seen also to enrich e.g. our anticipated combination of DSM [1] with the Axiomatic Design model [2], which structurally extends approaches dealing with the unstructured matrix view [3], [4]. The DSM view involves modularity of component, people and activity, whereas the Axiomatic Design model involves customer, function, product and production. The underlying logical scope appears in Section II, including the view of combination of models with respect to their granularity and modularity.

Respective logical enrichment of DSM and Axiomatic Design is prerequisite to a **logical merger of these two models into a unified conceptual framework for information and process for products and production**. Information basically resides within some use-case subprocess task, in turn being part of a larger process. Communication involving people performing activities is therefore not just a transmission but also requires transformation of data from a particular subprocess to be integrated and used within another subprocess. **We thereby iterate complexity of problems and integration of solutions**. Because of our logical model, industrial applications and required systems-oriented research are done in iteration, explained within a common interdisciplinary conceptual framework.

II. Logical Approach

A. Common Interdisciplinary Conceptual Framework

Smart `[systems]` basically means smart `[components-people-activities]` in combination and interaction, i.e., smartness of the multidomain in DSM. This builds upon smart `[components]`, smart [*people*] and smart [*products*], and this also explicitly initiates the explanation how smartness resides within a logical framework. Smartness thus embraces being correct and consistent about information. Further, we make a distinction between smart [*system of systems*] and [*smart system*] of systems. Various definitions appear in these contexts. Smart products are typically expected at least to involve monitoring, allow control, invite optimization and to be autonomous, each

capacity and capability building upon the preceding one, so that e.g. in order to have control capacity, a product must have monitoring capability.

Logically connecting activities and functioning, components and production, and enabling information structures to enrich availability, provide the foundations for describing the wide variety of smartness. In a concept like *time to failure* it is important to realize how *failure* needs to specified and enhanced by the underlying signature. Otherwise a failure may remain seen as an event, i.e., simply just a constant operator in the signature.

Maintenance strategy, preventive schedule (prevention guideline), predictive schedule (risk estimation), corrective (intervention), and inspection (monitoring and assessment scales), all appear in one form or another in the private and public sector. From stakeholder point of view, the private and public sectors are similar in that they **share the logic of information structures** even if the content is entirely different. The private and the public sectors, on the other hand, differ in that the private sector is a B2B matter, whereas the public sector is Triple Helix [5], including also a formal political dimension in university-industry-government relationships.

B. Extending the relational-logical view

We define a logical SoS involving dialogue of various form, interaction and integration, interfacing and transferral of information and structures, within and between subsystems. Further, system availability [3] and task scheduling [4] are important disciplinary aspects. The provision of a functioning standard will be extended to explain the potential use of a multi-valued logic enriched classification of functioning (MCFu) as related to a similarly enriched classification of failures (MCFa).

The logical enhancement of DSM is briefly outlined as follows. DSM uses a set X of *system elements* to create the set product $X \times X$ viewed as a *matrix*. Bivalent *interactions* between elements form a relation $R \subseteq X \times X$ sometimes viewed as a digraph. That bivalent relation can equivalently be represented as a function $\rho_R : X \times X \to \{0, 1\}$, so that aRb, i.e., $(a, b) \in R$ if and only if $\rho_R(a, b) = 1$. The set X is initially unstructured, i.e., no order or operations of any kind are imposed on X. Elements $x \in X$ therefore have no initial annotation or attribution of any kind. However, once elements are given names so as to represent components, granularity is intuitively recognized, and respective modelling approaches decide about levels of detail. A typical view is that *rolling up into lesser detail* [1] is accepted if significant information or insight isn't lost. We formalize this by viewing system elements as *terms over a signature* $\Sigma = (S, \Omega)$, where S is the structure of sorts (types) and Ω is the structure of operators. The set of system elements is then initially enriched to a set of terms $\mathsf{T}_\Omega X$, where X is a set of variables. As an example, an element $actuator$ as a point in set without structure is symbolically meaningless, whereas $actuator(f(x_1, \ldots, f(x_n))$ as term builds upon $aactuator$ and f as operators, and $x_1, \ldots, x_n$ as variables. Similarly, a temperature control could be a term $EATC(g(y, z))$. A many-valued interaction between the actuator and the temperature control could then be given as

$$\rho(actuator(f(x_1, \ldots, f(x_n)), EATC(g(y, z)))$$

where $\rho : \mathsf{T}_\Omega X \times \mathsf{T}_\Omega X \to \{no, weak, strong\}$ is a three-valued relation over $\mathsf{T}_\Omega X$, which unravels hidden information as compared to modelling using X only. This notation is based on category theory, where $\mathsf{T}_\Omega : \mathtt{Set} \to \mathtt{Set}$ is a *term functor* [6] that can be extended to a *monad*. The monad properties allow substitutions of expressions within a term to be composable, so that the substitutions $x = g(y, z)$ and $z = 22$ as composed and applied to $EATC(x)$ leads to the term $EATC(g(y, 22))$. This is obvious when we use relations over $\mathsf{T}_\Omega X$, but once we want to have various structured sets of terms, then we need *monad compositions* [6]. Powersets are typical examples, where system elements in a rolled up view are clustered into subsets of system elements, and that subset is given a new name as a new element in the rolled up DSM. This is the first steps in the logical modelling of SoS.

In a many-valued logical system for describing a DSM, components and subcomponents can be viewed in the same logical framework, e.g., with $piston(\ldots)$ and $crankshaft(\ldots)$ as parts of a $crank(\ldots, piston(\ldots), \ldots, crankshaft(\ldots), \ldots)$. In a bivalent view, this enrichment is still not dramatic. However, when adding the possibility that the term functor acts over various *Goguen categories* [7], $\mathsf{T}_\Omega : \mathtt{Set}(Q) \to \mathtt{Set}(Q)$, where Q is the algebraic structure for the selected view of many-valuedness, then many-valuedness can be invoked in a wide variety of ways in the DSM, Axiomatic Design and TRIZ models. In order for this paper to be more self-contained, we included some detail of the term functor in the Appendix.

Note that typing is prerequisite to modularity, i.e., the set of sorts in the signature, with related type constructors [6]. Granularity then comes from using a more elaborate structure of operators, ranging from the most coarse-granular situation where Ω contains only constants (of different type) to finer-granularity involving a wide range of operators with various arities, and operating over a rich structure of sorts and constructed types.

Rolling up into lesser detail will actually provide a name for a subset or cluster of system elements. It is then immediately important to note how symmetry and associativity, with trivial reflexivity, of a relation, means that we have an equivalence relation. These relational properties mean that X subdivides into equivalent classes. Again, in the bivalent case, this is apparent and straightforward, but when moving to the many-valued case, symmetry and associativity are then also many-valued, and the subdivision of $\mathsf{T}_\Omega X$ becomes a non-trivial matter. Another matrix approach example is the Pugh selection matrix, where the arithmetic calculation ignores the intuitive order among the design criteria, where e.g. ease and time of an implementation precedes cost of it. Here is yet another

example where many-valued approaches to types and operations enable modelling of order, i.e., how different criteria have precedence over other criteria, hidden in the matrix.

We further provide a logical enrichment of the matrix/relational models involved, so that relational types and strengths are further enhanced by **symmetries, associativities and granularities**. This brings in the mathematical disciplines of logic and algebra as a complement to numeric and stochastics. The **lativity**[1] **of systems** is a key feature where design precedes simulation, logic precedes design, statistics precedes analytics, subsupply precedes supply.

III. System-Of-Systems

A. System-of-systems definition

The definition of system of systems (SoS) is shown to be **logically extendable** over current definitions [8], which mostly describe types of systems of systems rather than their information content. In doing so, we are thereby closer to a system of systems views described as a collection of task-oriented or dedicated systems that pool their resources and capabilities together to create a new, more complex system, which offers **more functionality** and **performance** than simply the sum of the constituent systems [9].

The goal of a SoS architecture is to get **maximum value out of a large system**, comprised of smaller systems, by understanding how each of the smaller systems work, and developing industrially suitable interfaces based on industrial needs and usability studies. Since industrial applications are significantly diverse, and manufacturing industries communicate both within and across their boundaries, they are therefore complex and involve people with diverse backgrounds. Hence, complexity and cooperation are very important challenges, and robust cooperation and coordination between systems of humans and computing devices is crucial for progress in development of research results and in application development. Additionally, and perhaps most importantly, in particular applications, **implementation, verification and validation** activities must **attain high robustness and trustworthiness**, which requires underlying logical structure from complexity and facilitating cooperation as well as improved product and production development process data management.

B. The information and process view of SoS

Our language and concept, given the logic framework with MCFu and MCFa classifications, will support transformation and transferal of data between robots, humans and companies (Internet of Things 2.0), to optimally support the manufacturing industry in their contractor-subcontractor business relations, where sustainable production is measurable by key performance as described by our logical framework.

[1]'Lative' is related to motion, and more specifically, motion 'to' and 'from', so when terms appear in sentences, terms 'move into' sentence, and 'appear within' sentences. At the same time, sentences 'move away from' terms, and separates terms from sentences. In comparison, 'ablative' is motion 'away', and nominative is static.

Formal process modelling languages, like UML with its Behavioral Modeling, SysML and BPMN, from the Object Management Group (OMG), will enable formal information structures to be integrated with the process structures in manufacturing applications. The logic and ontology of information structures in industrial applications will be used for markup purposes. Upscaling of applications should follow an overall strategy as well as concrete suggestions and specific steps for the upscaling process and progress.

C. Application domains

For example, in forest products and mining industries, the general need is to ascertain optimal operation and utility of their subsystems, i.e., within the normal operation parameters. The goal is to maintain normal operation and optimal efficiency e.g. by avoiding unplanned stops and energy loss. The supplier-customer relation e.g. with respect to service and maintenance is important in particular from a SoS perspective.

Further, the automotive industry faces new SoS challenges e.g. due to the ambition to develop self-driving cars. Doing so brings developments much more outside the car itself, with needs to consider traffic and the environment as parts of the overall SoS. Customer experience, including safety and quality, also becomes extended with other aspects still not considered. From SoS point of view, traffic as a conglomerate of cars is composed differently as compared to a car being assembled from its components. Apart from industrial applications, also in health care there are several well-known systems-of-systems that can be logically treated similarly with respect to modelling of **information and process**. Ageing is a typical area where health and social care need to interact and become integrated in common care pathways. Falls prevention can be view as a specific example. Falls represent a major cause of burden and death in older adults [10]. Enhanced models of care pathways, with enriched data attached to targeted subsystems based on the herein presented logics framework, will enable monitoring of the overall care SoS performance, including macro-, meso- and micro- levels.

IV. State-Of-The-Art

Originality of our **logical approach** is two-fold. On the one hand, the use of classification of functioning, as clearly separated from but connected with taxonomies of products and classifications of structure, is an original approach in the manufacturing industry. Further, many-valuedness annotated with codes, and the way codes as well as structures of codes are many-valued, is a novelty not yet seen within manufacturing. On the other hand, and still from disciplinary research point of view, our approach to enable many-valuedness at all levels and modules within a logical machinery, is unique within the logic and in particular within the many-valued logic community. Traditional many-valuedness e.g. in form of fuzzy sets and fuzzy logic, and in particular as appearing within fuzzy control techniques, is an untyped and numeric based technique that further relies only on unstructured relations, like in Zadeh's

compositional rule of inference [11], and invokes many-valuedness only as far as truth values are concerned. Whereas statistics produce uncertainty quantification, **many-valued logic provides algebraic computing with uncertainties**.

Logic is a structure containing signatures and constructed terms and, and statements or sentences *latively* constructed based on terms. Similarly, sentences and conglomerates of sentences are fundamental for entailments, models and satisfactions, in turn part of axioms, theories and proof calculi. This lativity is suitably expressed in category theory using functors and monads, where constructions act over underlying categories in form of monoidal categories. Category theory is thus a suitable metalanguage for logic, in particular when applications and typing of information must be considered. Uncertainty may reside in generalized powerset functors, and may be internalized in underlying categories. In both cases, suitable algebras must motor this uncertainty representation, and quantales are very suitable in this context. [20]

From a systems-oriented research point of view, logical enrichment of **availability simulation, faults and functioning**, DSM, Axiomatic Design, TRIZ, and other information and process related models, connect not only to methods in modern logic, but also to the historical traditions in logic such as represented by sets and types in Principia Mathematica or type theory as initiated by Schönfinkel's Bausteine [12], Curry's functionality [13] and Church's simple typing [14]. Göttingen and Hilbert's foundations of mathematics were a driving force for Gödel and his work in Vienna, and many-valued logic was started by the Lwow-Warsaw school. Kolmogorov's Aufgabe [15], related to the TRIZ view of inventiveness, as a task rather than a problem is a broader foundation for algorithm as later developed by Turing [16].

For classification purposes, type constructors must make use of category in order to enable underlying categories that represent uncertainties, using a three-level signature [6], where structured powerset constructors can be handled properly. The generic scale of uncertainties resides in those underlying categories, and the algebraic use of the scale [19] is a technique that is unavailable in traditional type theory approaches e.g. within homotopy type theory (HoTT) [17]. A typical first step to many-valuedness is adding an unspecified to two-valuedness.

Relations like trees and matrices are basically unstructured sets and relations, which e.g. means that our approach potentially enriches design methodologies which basically use matrix computations to relate physical product structure with function. Values in matrices are always just numerical values, and also **detached** from any nomenclatures. Logic enables representation where relations become enriched with more structure and attached with classifications. Enrichment of structure for products then spills over to enrichment of structures in production. Thus, engineers are potentially provided with tools that enables improved quantification for design evaluation as well as for quality assurance and control, e.g. as those appearing in failure mode and effects analysis (FMEA) [18].

Industrial product development challenges relate, on the one hand, to information structures, respectively, for customer, function, product and production, and, on the other hand, to compatibility and transformations between these structures. DSM is typically used for internal structure descriptions, whereas Axiomatic Design is typically used for the transformations. See [1], [2] for details and examples. In both models, matrices only, with numerical and untyped values, are used to represent relational structures. Transformation of information within these models should indeed not just restrict to mapping of matrix content in an unstructured manner, but rather start from identifying the relational content, and thereby enable expansion and enrichment towards using structure preserving transformations.

A. DSM

In its most rudimentary form, a Design Structure Matrix [1] is a relation $\rho : X \times X \to \{no, yes\}$, where X is a set of *system elements*. The engineering understanding of such a system element may be very complex, but in the mathematical model of it, it is just an element, or actually a name of an element. A matrix is more appealing if yes values appear close to the diagonal, which implies a certain **nearness** between system elements. With values distant from the diagonal, clustering can provide conglomeration points with values that make the clustered matrix appear more diagonal.

The bivalent interaction is sometimes extended to a three-valued interaction with names in the three-valued set of truths being 'strong interaction', 'weak interaction' and 'no interaction'. There are no logical connectives, so there is no explicit propositional logic annotated with DSM. In the traditional DSM model there is also no considerations for a 'not specified', which would invite to viewing that three-valued truth set as a non-commutative quantale [20] or commutative Bocvar-Kleene algebra [21], [22]. Kleene called than in-between value 'unspecified', whereas Bocvar called it 'senseless'. Kleene used his three-valued logic e.g. to model partial functions within recursion, so 'unspecified' in logical connection with something specified is an interchangeable (commutative) operation.

In addition to bivalent interaction, multivalent interaction exists also as related to the use of typed interactions, like the one with four types, respectively, for 'spatial', 'energy', 'information' and 'materials'. Each type is valuated within a 5-scale $\{-2, -1, 0, +1, +2\}$, with -2 for 'detrimental', and $+2$ for 'required'. The internal algebraic structure of the set of components is, however, not given.

Product, organization and process structures, respectively, involving components, people and activities, also appear as integrated in a multidomain architecture.

B. Axiomatic Design

Information and process development addresses the challenge to combine components and product taxonomies with activity and process hierarchies, in order to support decision-makers, engineers and customers. Many-valued logic enriches

the anticipated combination of DSM with the Axiomatic Design model [2], which in effect goes far beyond approaches dealing only with the unstructured matrix view. Whereas the DSM view involves modularity of component, people and activity, the Axiomatic Design model involves respective domains for customer, function, (physical) product and (observable) process.

Decoupled design, in case of a triangular matrix, is desirable, as this enables sequential consideration within the domains. Axiomatic Design theory is product design which start from using customer attributes (CA) as a basis for functional requirements (FR), in turn to map over to design parameters (DP), and from there arriving at process variables (PV). The FR to DP mapping is the most critical one.

Attributes, requirements, parameters and variables are all identified by names only, i.e., they are logically constants (0-ary operators). Independence and Information Axioms are formulated using these constants, and decomposition within and between (zigzagging) domains is simply saying that a constant becomes the name of a set of new constants, building up a hierarchy of constants.

Axiomatic Design involves a matrix view between domains, but not within a domain as in the case of DSM. Integrating DSM into Axiomatic Design using the unstructured and traditional models is suggested in [23]. Integration based on our structured approach will additionally involve structure-preservation.

Respective logical enrichment of DSM and Axiomatic Design is prerequisite to a logical merger of these two models into a unified conceptual framework for information and process for products and production. Information basically resides within some use-case subprocess task, in turn being part of a larger process. Communication involving people performing activities is therefore not just a transmission but also requires transformation of data from a particular subprocess to be integrated and used within another subprocess. We can thereby iterate complexity of problems and integration of solutions.

C. TRIZ

TRIZ [24] as a *theory of inventive problem solving* is a general model, and an informal model as more formal data, logical or computational models are not included. For algorithms in ARIZ [25], data models are required, but are still on a very general level. Objects and classes, like those modeled e.g. by Class Diagrams in UML, can be used, but will not suffice in order to represent and solve contradictions in TRIZ' contradiction matrix, since TRIZ inherently involves processes and behavior. The UML Class Diagram is simply a data model. Furthermore, the features in the contradiction matrix are of a wide variety of types, which are lumped together into one "set of features". This makes it unsuitable to view the matrix as a basis for a relational model. The set of features must be dissected and structured, and the features themselves must be enriched and further specified.

The 40 TRIZ Principles are also very different in nature and content, and are more like principle of common sense than principles of reasoning. The 'Preliminary action' principle for the related features 'Reliability' and 'Loss of time' basically recommends prevention of faults, prior to detection of faults, should they happen. Improving feature 'Ease of repair' as related to worsening feature 'Device complexity' involves the 'Segmentation' principle with respect to the need for increasing transparency and modularity. In logic we would say that "this logic is sound but not complete" is a metalogical statement, whereas "if you have to use a logic that sound but not complete, try to use a logic that is as complete as possible" is a principle more in the style of common sense.

Several features in the contradiction matrix are physical and/or geometrical, which logically can be expressed by structured terms rather than as unstructured concepts, as typically seen within description logic. However, description logic can be enriched [26] in order to make it better fit as a logic for TRIZ like ontologies. Features like 'measurement accuracy' and 'manufacturing precision' can be managed by adopting a logical framework with underlying signatures acting over a monoidal category [27] representing multivalence and uncertainty. Features like 'loss of ...' and 'reliability' can be logically detailed for particular problem contexts, where 'loss of ...' features are related to functioning rather than faults and failures. The 'Ease of ...' features can only be described by subprocesses, so in the case of UML we would need to consider to use Behavior Diagrams, or, if staying within OMG standards, move to using SysML. The 'Productivity' feature is closer to being analyzed within the BPMN model. Examples with applications in crisis management have been developed in [28].

V. UPSCALING

Industrial upscaling is more than just replication or multi-piloting, i.e., not just going from a small pilot to a large pilot. Specific pilots and cases expand within and across companies, and this expansion requires availability and acceptance of the common language, so that when scaling up, geographically distributed teams and cooperating stakeholders realize the need to be even more precise about underlying logical-relational information structures as compared to just doing specific and local pilots. Thus, the logical approach presented here supports improved upscaling.

A. Upscaling strategy

Upscaling should follow an overall strategy as well as concrete suggestions and specific steps for the upscaling process and progress. Guidelines are required in order to manage the framework that involves the critical elements in scaling up. Logic as an ingredient in manufacturing has its upscaling focus (at least) on the following:

- Design and manufacturing issues, which are supported by company specific and locally generated well proven practices of systematic effectiveness and feasibility, are key factors that help to increase the likelihood for successful and sustained upscaling.

- The balance between the rigid scaling up models and the pilot implementations residing within that scaling up process and structure needs to be maintained.
- Scaling up may often require additional managerial and financial input, and also an acceptance that upscaling usually implies a longer timeframe as compared to typically seen in ordinary company project cycles.

B. Steps in upscaling

A general upscaling model, building upon our logical-relational machinery, includes the following steps:

i. Descriptive identification of good practices using our language and classification, and enabling refined and broadened data collections and monitoring.
ii. Prescriptive analytics (numerics) and assessment (logics) of the viability and benefits of upscaling within domains of industries.
iii. Logical classification of good practices for replication, whenever feasible, and for suitably adapted implementation with respect to a variety of competences and industrial circumstances.
vi. Facilitation of partnership for scaling up within and across corporations, making resources available.
v. Identify key success factors for generalized implementation, and recognize lessons learnt.

Compliance within upscaling and reinforced changes in working pattern is strongly emphasized, where also personal skill and personalized behaviour is fundamental. Upscaling should build upon excellence achieved within information structuring cultures in the production industry, and aims at enriching these structures using the logical-relational approach.

VI. Conclusions

We have shown how to enrich the language used in the manufacturing industry regarding information structure and its representation for products and production processes. This is achieved using many-valued logic, in order to complement the numerical approach commonly appearing in such representations. We underline the importance of utilizing mathematical disciplines like algebra and logic side-by-side with the utility of analysis and stochastics. Within a communicating resources perspective, many-valued logic will potentially contribute to *information structuring* and *management of data complexity* in big data applications.

In future papers we will further introduce a structure for functioning classification in order to complement and interact with the traditional view of faults and failures in product and production subsystems. Our concept of logic modelling is based on nomenclatures, classification and ontology as information structures part of and supporting production, and indeed in production where information and process is tightly connected. Support for manufacturing, machine simulation, and modelling, then becomes part of design processes, e.g. in cases like cars and car production, as related to subsystems such as powertrain, electronics or interior,

ACKNOWLEDGEMENT

This work is supported by the *Logic in Manufacturing* (LiM) project, with gratefully acknowledged funding from the Swedish Innovation Agency (VINNOVA) PRODUKTION2030 programme.

References

[1] S. D. Eppinger, T. R. Browning, *Engineering Systems: Design Matrix Methods and Applications*, MIT Press, 2012.
[2] N. P. Suh, *Axiomatic Design: Advances and Applications*, Oxford University Press, 2001.
[3] M. Löfstrand, M. Karlberg, J. Andrews, L. Karlsson, *Functional product system availability: simulation driven design and operation through coupled multi-objective optimization*, International Journal of Product Development **13** (2011), 119-131.
[4] S. Reed, J. Andrews, S. Dunnett, P. Kyösti, B. Backe, M. Löfstrand, L. Karlsson, *A modelling language for maintenance task scheduling*, In: 11th International PSAMS and ESREL 2012 Conference. vol. 1, 201-211.
[5] H. Etzkowitz, L. Leydesdorff, *The Triple Helix—University-Industry-Government Relations: A Laboratory for Knowledge-Based Economic Development*, EASST Review **14** (1995), 14-19.
[6] P. Eklund, M.A. Galán, R. Helgesson, J. Kortelainen, *Fuzzy terms*, Fuzzy Sets and Systems **256** (2014), 211-235.
[7] P. Eklund, J. Kortelainen, L. N. Stout, *Adding fuzziness using a monadic approach to terms and powerobjects*, Fuzzy Sets and Systems **192** (2012), 104-122.
[8] M. Jamshidi, *System-of-Systems Engineering - A Definition*, IEEE SMC 2005, 10-12 Oct. 2005.
[9] S. Popper, S. Bankes, R. Callaway, D. DeLaurentis, *System-of-Systems Symposium: Report on a Summer Conversation*, July 21âĂ¿22, 2004, Potomac Institute for Policy Studies, Arlington, VA.
[10] H. Blain, F. Abecassis, P. Adnet, B. AlomÃĺne, M. Amouyal, B. Bardy, et al., *Living Lab Falls-MACVIA-LR: The falls prevention initiative of the European Innovation Partnership on Active and Healthy Ageing (EIP on AHA) in Languedoc Roussillon*, Eur Geriatr Med. **5** (2014), 416-425.
[11] L. Zadeh, *Outline of a new approach to the analysis of complex systems and decision processes*, IEEE Trans. Systems, Man and Cybernetics **3** (1973), 28âĂ¿44.
[12] M. Schönfinkel, *Über die Bausteine der mathematischen Logik*, Mathematische Annalen **92** (1924), 305-316.
[13] H. B. Curry, *Functionality in combinatory logic*, Proc Natl Acad Sci USA **20** (1934), 584-590.
[14] A. Church, *A formulation of the simple theory of types*, The journal of symbolic logic **5** (1940), 56-68.
[15] A. N. Kolmogorov, Zur Deutung der intuitionistischen Logik, Mathematische Zeitschrift **35** (1932), 58-65.
[16] A. M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society. Ser. 2, Vol. 42 (1937), 230âĂ¿65.
[17] *Homotopy Type Theory: Univalent Foundations of Mathematics*, The Univalent Foundations Program, Institute for Advanced Study, 2013.
[18] *MIL-P-1629 - Procedures for performing a failure mode effect and critical analysis*, United States Department of Defense (9 November 1949).
[19] P. Eklund, U. Höhle, J. Kortelainen, *A Survey on the categorical term construction with applications*, Fuzzy Sets and Systems, Available online 13 July 2015.
[20] P. Eklund, U. Höhle, J. Kortelainen, *Non-commutative quantales for many-valuedness in applications*, Proc. IPMU 2016, to appear.
[21] D. A. Bocvar, *Ob odnom trechznacnom iscislenii i ego primenenii k analizu paradoksov klassiceskogo funkcional'nogo iscislenija*, Mat. Sbornik **4** (1938), 287-308.
[22] S. C. Kleene, *On notation for ordinal numbers*, J. Symbolic Logic **3** (1938), 150-155.
[23] D. Tang, R. Zhu, S. Dai, G. Zhang, *Enhancing axiomatic design with design structure matrix*, Concurrent Engineering: Research and Applications **17** (2009), 129-137.
[24] G. S. Altshuller, *Creativity as an Exact Science*, Gordon & Breach, 1984.

[25] G. S Altshuller, *The Innovation Algorithm: TRIZ, systematic innovation and technical creativity*, Technical Innovation Center, Worcester, MA., 1999.
[26] P. Eklund, *The syntax of relations*, Proc. IPMU 2016, to appear.
[27] S. Mac Lane, *Categories for the Working Mathematician*, Graduate Texts in Mathematics **5** (second ed.), Springer, 1998.
[28] P. Eklund, M. Johansson, J. Karlsson, R. Åström, *BPMN and its Semantics for Information Management in Emergency Care*, Fourth 2009 International Conference on Convergence and Hybrid Information Technology (ICCIT 2009), IEEE Computer Society, 273-278.

APPENDIX

In the following we briefly introduce notation and constructions needed in our descriptions related to our SoS related logic, and in particular for its underlying signatures and terms. The many-sorted term monad $\mathbf{T}_\Sigma$ over $\mathtt{Set}_S$, the many-sorted category of sets and functions, where $\Sigma = (S, \Omega)$ is a signature, can briefly be described as follows. For a sort (i.e. type) $\mathtt{s} \in S$, we have sort specific functors $\mathsf{T}_{\Sigma,\mathtt{s}} : \mathtt{Set}_S \to \mathtt{Set}$, so that

$$\mathsf{T}_\Sigma(X_\mathtt{s})_{\mathtt{s}\in S} = (\mathsf{T}_{\Sigma,\mathtt{s}}(X_\mathtt{s})_{\mathtt{s}\in S})_{\mathtt{s}\in S}.$$

The important recursive step in the term construction is

$$\mathsf{T}^\iota_{\Sigma,\mathtt{s}}(X_\mathtt{s})_{\mathtt{s}\in S} =$$

$$\coprod_{\mathtt{s}_1,\ldots,\mathtt{s}_m} (\Omega^{\mathtt{s}_1\times\cdots\times\mathtt{s}_m\to\mathtt{s}})_{\mathtt{Set}_S} \times \mathsf{arg}^{\mathtt{s}_1\times\cdots\times\mathtt{s}_m} \circ \bigcup_{\kappa<\iota} \mathsf{T}^\kappa_\Sigma(X_\mathtt{s})_{\mathtt{s}\in S}$$

and then with

$$\mathsf{T}^\iota_\Sigma(X_\mathtt{s})_{\mathtt{s}\in S} = (\mathsf{T}^\iota_{\Sigma,\mathtt{s}}(X_\mathtt{s})_{\mathtt{s}\in S})_{\mathtt{s}\in S},$$

we finally arrive at the term functor

$$\mathsf{T}_\Sigma = \bigcup_{\iota<\bar{k}} \mathsf{T}^\iota_\Sigma.$$

The term functor construction can be extended so that $\mathsf{T}_\Sigma : \mathtt{C} \to \mathtt{C}$ operates more generally over monoidal biclosed categories $\mathtt{C}$. If $\mathtt{C}$ is $\mathtt{Set}$, we have the construction above, and with the Goguen category $\mathtt{Set}(Q)$, where Q is a quantale, we have a multivalent and typed situation enabled by the signature acting over the selected underlying category.

Term functors constructed in this way can be extended to monads, so that substitution can be composed. A substitution in this categorical context is a morphism in the Kleisli category of the related monad.

Monad compositions further enable to arrive at generalized sets of terms, where the typical example is composing the term functor T_Σ with the powerset functor P in order to obtain the monad $\mathbf{P} \circ \mathbf{T}_\Sigma$. More elaborate generalized set functors Φ can be applied in order to make use of the composition $\mathbf{\Phi} \circ \mathbf{T}_\Sigma$.

For more detail, and for the purely categorical constructions of the corresponding term monads, the reader is referred to [6].

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 19–23
DOI: 10.15439/2016F564
ACSIS, Vol. 9. ISSN 2300-5963

Computing the minimal solutions of finite fuzzy relation equations on linear carriers

Juan Carlos Díaz-Moreno, Jesús Medina, Esko Turunen
Department of Mathematics
University of Cádiz, Spain
Email: {juancarlos.diaz,jesus.medina}@uca.es
Research Unit Computational Logic.
Vienna University of Technology. Wien, Austria
Email: esko.turunen@tut.fi

***Abstract*—Fuzzy relation equations is an important tool for managing and modeling uncertain or imprecise datasets, which has useful applied to, e.g. approximate reasoning, time series forecast, decision making, fuzzy control, etc. This paper considers a general fuzzy relation equation, which has minimal solutions, if it is solvable. In this case, an algebraic characterization is introduced which provides an interesting method to compute minimal solutions in this general setting.**

I. Introduction

FUZZY relation equations were introduced by E. Sanchez in the seventies [11]. These equations have widely been studied in different papers [1], [3], [6]. For example, they have proven that the set of solutions of solvable fuzzy relation equations is a upper-preserving complete lattice in which the greatest solutions is completely determined. Nevertheless, the computation of minimal solutions is not so direct. These solutions have also been studied in several papers [2], [4], [10], [12], [17], [15], [16], [18], [13] and several algorithms have been developed, but in restrictive frameworks, restrictions that limit the flexibility of the possible applications.

Hence, first of all, it is fundamental to study general frameworks in which the minimal solutions of each solvable fuzzy relation equation exist and that each solution will be between the greatest solution and a minimal solution.

This paper considers a general setting, in which the operators may neither be commutative nor associative and they only need to be monotone and residuated inf-preserving mappings of non-empty sets on the right argument. The linearity of the carrier, together with the inf-preserving property, ensures the existence of minimal solutions whenever a solution exists.

Mainly, this paper introduces a procedure in order to obtain the minimal solutions of a solvable of the introduced general fuzzy relation equations. Moreover, we have presented a detailed algorithm to compute these important solutions, together with several illustrative examples.

II. General fuzzy relation equations

Throughout this paper we will consider a complete linear lattice $(L, \preceq)$, in which the bottom and the top elements exist and they are denoted as $0, 1$, respectively. Given a set V, the ordering $\preceq$ in the lattice induces a partial order on the set of L-fuzzy subsets of V, L^V. This ordering provides to L^V the structure of a complete lattice.

A general residuated operator will also be used in this paper to define the fuzzy relation equation, as in [8]. This residuated operator will be denoted as $\odot\colon L \times L \to L$, which is order preserving in both arguments and there exists another operator $\to\colon L \times L \to L$, satisfying the following adjoint property with the conjunctor $\odot$

$$x \odot y \preceq z \quad \text{if and only if} \quad y \preceq x \to z \tag{1}$$

for each $x, y, z \in L$. This property is equivalent to say that $\odot$ preserves supremums in the second argument; $x \odot \bigvee\{y \mid y \in Y\} = \bigvee\{x \odot y \mid y \in Y\}$, for all $Y \subseteq L$.

These operators, as were noted in [8], generalize other kind of residuated pairs [7], [5], since only the monotonicity and the adjoint property are considered.

Definition 1. *Given the pair $(\odot, \to)$, a* fuzzy relation equation *is the equation:*

$$R \circ X = T, \tag{2}$$

where $R\colon U \times V \to L$, $T\colon U \times W \to L$ are given finite L-fuzzy relations and $X\colon V \times W \to L$ is unknown; and $R \circ X\colon U \times W \to L$ is defined, for each $u \in U$, $w \in W$, as

$$(R \circ X)\langle u, w\rangle = \bigvee\{R\langle u, v\rangle \odot X\langle v, w\rangle \mid v \in V\}.$$

It is well known that the fuzzy relation equation (2) has a solution if and only if

$$(R \Rightarrow T)\langle v, w\rangle = \bigwedge\{R\langle u, v\rangle \to T\langle u, w\rangle \mid u \in U\}$$

is a solution and, in that case, it is the greatest solution, see [7], [11], [14].

III. Computing minimal solutions on linear lattices

Definition 2. *Given an operator $\odot\colon L \times L \to L$, we will say that it holds the* IPNE-condition *(making reference to that $\odot$ is Infimum Preserving of arbitrary Non-Empty sets), if it verify*

$$a \odot \bigwedge B = \bigwedge\{a \odot b \mid b \in B\} \tag{3}$$

for each element $a \in L$ and each non-empty subset $B \subseteq L$.

From now on, let us consider a general solvable fuzzy relation equation (2), where R, X, T are finite and $\odot$ satisfies the IPNE-condition.

First of all, the auxiliary sets V_{uw} need to be introduced, which are associated with the elements $u \in U$, $w \in W$ and the greatest solution $R \Rightarrow T$. Since for each $u \in U$, $w \in W$

$$\bigvee\{R\langle u, v\rangle \odot (R \Rightarrow T)\langle v, w\rangle \mid v \in V\} = T\langle u, w\rangle, \qquad (4)$$

L is linear and V is finite, there exists at least one $v_s \in V$ validating the equation

$$R\langle u, v_s\rangle \odot (R \Rightarrow T)\langle v_s, w\rangle = T\langle u, w\rangle. \qquad (5)$$

Therefore, the set

$$V_{uw} = \{v \in V \mid R\langle u, v\rangle \odot (R \Rightarrow T)\langle v, w\rangle = T\langle u, w\rangle\}$$

is not empty and, for all $v \notin V_{uw}$, the strict inequality $R\langle u, v\rangle \odot (R \Rightarrow T)\langle v, w\rangle < T\langle u, w\rangle$ holds.

Each v_s in V_{uw} will provide a fuzzy subset S_{uws} as follows: Given $v_s \in V_{uw}$, we have that

$$\{d \in L \mid R\langle u, v_s\rangle \odot d = T\langle u, w\rangle\} \neq \varnothing$$

and the infimum $\bigwedge\{d \in L \mid R\langle u, v_s\rangle \odot d = T\langle u, w\rangle\} = e_s$ also satisfies the equality

$$R\langle u, v_s\rangle \odot e_s = T\langle u, w\rangle$$

by the IPNE-condition. These elements are used to define the fuzzy subsets of V, $Z_{uws}\colon V \to L$, defined by

$$Z_{uws}(v) = \begin{cases} e_s & \text{if } v = v_s \\ 0 & \text{otherwise} \end{cases}$$

which form the set Z_{uw}, that is $Z_{uw} = \{Z_{uws} \mid v_s \in V_{uw}\}$, for each $u \in U$, $w \in W$. These sets will be used to characterize the set of solutions of Equation (2) by the notion of *covering*.

Theorem 3. *The L-fuzzy relation $X\colon V \times W \to L$ is a solution of a solvable Equation (2) if and only if $X \preceq (R \Rightarrow T)$ and, for each $w \in W$, the fuzzy subset $X_w\colon V \to L$, defined by $X_w(v) = X\langle v, w\rangle$, is a cover of $\{Z_{uw} \mid u \in U\}$.*

As a consequence, the minimal solutions are characterized by the minimal covers.

Corollary 4. *$X\colon V \times W \to L$ is a minimal solution of Equation (2) if and only if, for each $w \in W$, $X_w\colon V \to L$, defined by $X_w(v) = X\langle v, w\rangle$, is a minimal cover of $\{Z_{uw} \mid u \in U\}$.*

Hence, from the corollary above, minimal solutions of the fuzzy relation equation (2) are obtained from $R \Rightarrow T$. Next, the detailed algorithms are introduced.

Module *MINIMAL_COVERING* uses an usual algorithm in order to compute minimal covering of subsets.

Example III.1. *Let us assume the standard MV–algebra [9], that is, $L = [0, 1]$ is the unit interval, $\odot\colon L \times L \to L$ is the Łukasiewicz operator defined by $x \odot y = \max\{0, x + y - 1\}$ and $\to\colon L \times L \to L$ its residuated implication, defined by $y \to z = \min\{1, 1 - y + z\}$, for all $x, y, z \in [0, 1]$.*

input : Universes U, V and W, the fuzzy relations $R\colon U \times V \to L$ and $T\colon U \times W \to L$
output: MSS= Set of minimal solutions of the fuzzy relation equation $R \circ X = T$

```
MSS := [ ];
S := R ⇒ T, which is the greatest solution of R ∘ X = T ;
for k ← 1 to |U| and j ← 1 to |W| do
    Z_kj := [ ];
    for i ← 1 to |V| do
        Z_kji:= zeros row of |V|-order;
        if R[k,i] ⊙ S[i,j] = T[k,j] then
            e_i := min{y ∈ [0,1] | R[k,i] &_L y = T[k,j]};
            update Z_kji[i] by the value e_i;
            add Z_kji to the list Z_kj;
        end
    end
end
for j ← 1 to |W| do
    Z_j := [Z_1j, ..., Z_|U|j];
    [X_j^1, ..., X_j^{l_j}] :=MINIMAL_COVERING(Z_j);
end
for h_1 ← 1 to l_1 and ... h_|W| ← 1 to l_|W| do
    X_{h_1...h_|W|} := zeros matrix of |V| × |W|-order;
    for j ← 1 to |W| and i ← 1 to |V| do
        X_{h_1...h_|W|}[i,j] := X_j^{h_j}[i] ;
    end
    add X_{h_1...h_|W|} to the list MSS;
end
```

Algorithm 1: PMINSOLUTIONS(R, T)

Given $U = \{u_1, u_2, u_3\}$, $V = \{v_1, v_2, v_3\}$ $W = \{w_1, w_2, w_3\}$ and the fuzzy relation equations, defined from the following tables

R	v_1	v_2	v_3
u_1	0.9	0.5	0.9
u_2	0.2	0.9	0.7
u_3	0.8	0.6	0.9

and

T	w_1	w_2	w_3
u_1	0.8	0.4	0.7
u_2	0.6	0.7	0.3
u_3	0.8	0.4	0.6

direct computation shows that the relation $R \Rightarrow T$, defined from the table

$R \Rightarrow T$	w_1	w_2	w_3
v_1	0.9	0.5	0.8
v_2	0.7	0.8	0.4
v_3	0.9	0.5	0.6

is the greatest solution of Equation (2). During the verification we go through the following calculations:
When computing $(R \circ (R \Rightarrow T))\langle u_1, w_1\rangle = 0.8$, we consider the maximum of

$$\begin{aligned} R\langle u_1, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_1\rangle &= 0.9 + 0.9 - 1 = 0.8 \\ R\langle u_1, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_1\rangle &= 0.5 + 0.7 - 1 = 0.2 \\ R\langle u_1, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_1\rangle &= 0.9 + 0.9 - 1 = 0.8 \end{aligned}$$

Notice that from v_1 and v_3 we get the maximum. Hence, in order to obtain this maximum, we only need to consider $\{v_1\}$ or $\{v_3\}$. Moreover, the values 0.9 associated with v_1 and 0.9 associated with v_3 cannot be decreased because, if we decrease them, a value less than 0.8 will be obtained in the computation and we do not reach a solution. Therefore, the first column of a solution of Equation (2) could be any column in the set:

$$Z_{1,1} = \{\begin{pmatrix}0.9\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0.9\end{pmatrix}\}$$

However, we need to verify that the other two equalities also hold. Consequently, the equality $(R \circ (R \Rightarrow T))\langle u_2, w_1\rangle = 0.6$ is studied similarly to the previous procedure. The value $(R \circ (R \Rightarrow T))\langle u_2, w_1\rangle$ is the maximum of the values

$$\begin{array}{lcccl} R\langle u_2, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_1\rangle & = & 0.2 + 0.9 - 1 & = & 0.1 \\ R\langle u_2, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_1\rangle & = & 0.9 + 0.7 - 1 & = & 0.6 \\ R\langle u_2, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_1\rangle & = & 0.7 + 0.9 - 1 & = & 0.6 \end{array}$$

for which $\{v_2\}$ or $\{v_3\}$ is only necessary and so, the first column of a solution of Equation (2) could be one element of the set:

$$Z_{2,1} = \{\begin{pmatrix}0\\0.7\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0.9\end{pmatrix}\}$$

Finally, when computing $(R \circ (R \Rightarrow T))\langle u_3, w_1\rangle = 0.8$ we pass by

$$\begin{array}{lcccl} R\langle u_3, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_1\rangle & = & 0.8 + 0.9 - 1 & = & 0.7 \\ R\langle u_3, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_1\rangle & = & 0.6 + 0.7 - 1 & = & 0.3 \\ R\langle u_3, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_1\rangle & = & 0.9 + 0.9 - 1 & = & 0.8 \end{array}$$

In this case, only v_3 is necessary and one column is only considered:

$$Z_{3,1} = \{\begin{pmatrix}0\\0\\0.9\end{pmatrix}\}$$

We observe that

$$K = \begin{pmatrix}0\\0\\0.9\end{pmatrix} \in Z_{1,1} \cap Z_{2,1} \cap Z_{3,1},$$

so K is the only minimal column which, in an intuitive sense, covers the set $Z_1 = \{Z_{1,1}, Z_{2,1}, Z_{3,1}\}$. Moreover, we conclude that a fuzzy relation X_1, defined as

X_1	w_1	w_2	w_3
v_1	0	0.5	0.8
v_2	0	0.8	0.4
v_3	0.9	0.5	0.6

solves the fuzzy relation equation (2).

Next, we consider the second column of $R \Rightarrow T$, which provides a different case. For $(R \circ (R \Rightarrow T))\langle u_1, w_2\rangle = 0.4$ we have

$$\begin{array}{lcccl} R\langle u_1, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_2\rangle & = & 0.9 + 0.5 - 1 & = & 0.4 \\ R\langle u_1, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_2\rangle & = & 0.5 + 0.8 - 1 & = & 0.3 \\ R\langle u_1, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_2\rangle & = & 0.9 + 0.5 - 1 & = & 0.4 \end{array}$$

Hence, the maximum is obtained from v_1 or v_3 and, therefore, the following set is considered:

$$Z_{1,2} = \{\begin{pmatrix}0.5\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0.5\end{pmatrix}\}$$

For $(R \circ (R \Rightarrow T))\langle u_2, w_2\rangle = 0.7$ we have

$$\begin{array}{lcccl} R\langle u_2, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_2\rangle & = & 0 & & \\ R\langle u_2, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_2\rangle & = & 0.9 + 0.8 - 1 & = & 0.7 \\ R\langle u_2, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_2\rangle & = & 0.7 + 0.5 - 1 & = & 0.2 \end{array}$$

Consequently, the subset obtained is

$$Z_{2,2} = \{\begin{pmatrix}0\\0.8\\0\end{pmatrix}\}$$

For $(R \circ (R \Rightarrow T))\langle u_3, w_2\rangle = 0.4$ we have

$$\begin{array}{lcccl} R\langle u_3, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_2\rangle & = & 0.8 + 0.5 - 1 & = & 0.3 \\ R\langle u_3, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_2\rangle & = & 0.6 + 0.8 - 1 & = & 0.4 \\ R\langle u_3, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_2\rangle & = & 0.9 + 0.5 - 1 & = & 0.4 \end{array}$$

Hence, the assumed subset of columns is

$$Z_{3,2} = \{\begin{pmatrix}0\\0\\0.5\end{pmatrix}, \begin{pmatrix}0\\0.8\\0\end{pmatrix}\}$$

In this case, we observe that $Z_{1,2} \cap Z_{2,2} \cap Z_{3,2} = \varnothing$. However,

$$\begin{pmatrix}0\\0\\0.5\end{pmatrix}, \begin{pmatrix}0\\0.8\\0\end{pmatrix} \leq \begin{pmatrix}0\\0.8\\0.5\end{pmatrix} = \begin{pmatrix}0\\0.8\\0\end{pmatrix} \vee \begin{pmatrix}0\\0\\0.5\end{pmatrix},$$

$$\begin{pmatrix}0.5\\0\\0.\end{pmatrix}, \begin{pmatrix}0\\0.8\\0\end{pmatrix} \leq \begin{pmatrix}0.5\\0.8\\0\end{pmatrix} = \begin{pmatrix}0\\0.8\\0\end{pmatrix} \vee \begin{pmatrix}0.5\\0\\0\end{pmatrix}$$

and $\begin{pmatrix}0\\0.8\\0.5\end{pmatrix}, \begin{pmatrix}0.5\\0.8\\0\end{pmatrix}$ are the only minimal columns which, again in an intuitive sense, cover the set $Z_2 = \{Z_{1,2}, Z_{2,2}, Z_{3,2}\}$. Moreover, we conclude that the fuzzy relations X_2 and X_3, defined as

X_2	w_1	w_2	w_3
v_1	0	0	0.8
v_2	0	0.8	0.4
v_3	0.9	0.5	0.6

X_3	w_1	w_2	w_3
v_1	0	0.5	0.8
v_2	0	0.8	0.4
v_3	0.9	0	0.6

solve the fuzzy relation (2). Finally, the values in the third column of $R \Rightarrow T$ are reduced.

For $(R \circ (R \Rightarrow T))\langle u_1, w_3\rangle = 0.7$, we compute

$$\begin{array}{lcccl} R\langle u_1, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_3\rangle & = & 0.9 + 0.8 - 1 & = & 0.7 \\ R\langle u_1, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_3\rangle & = & 0 & & \\ R\langle u_1, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_3\rangle & = & 0.9 + 0.6 - 1 & = & 0.5 \end{array}$$

Hence, $Z_{1,3} = \{\begin{pmatrix}0.8\\0\\0\end{pmatrix}\}$.

For $(R \circ (R \Rightarrow T))\langle u_2, w_3\rangle = 0.3$*, we have*

$$\begin{array}{lcccl} R\langle u_2, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_3\rangle & = & 0.2 + 0.8 - 1 & = & 0 \\ R\langle u_2, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_3\rangle & = & 0.9 + 0.4 - 1 & = & 0.3 \\ R\langle u_2, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_3\rangle & = & 0.7 + 0.6 - 1 & = & 0.3 \end{array}$$

two possibilities providing two columns: $Z_{2,3} = \{\begin{pmatrix}0\\0.4\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0.6\end{pmatrix}\}$.

For $(R \circ (R \Rightarrow T))\langle u_3, w_3\rangle = 0.6$ *we have*

$$\begin{array}{lcccl} R\langle u_3, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w_3\rangle & = & 0.8 + 0.8 - 1 & = & 0.6 \\ R\langle u_3, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w_3\rangle & = & 0.6 + 0.4 - 1 & = & 0 \\ R\langle u_3, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w_3\rangle & = & 0.9 + 0.6 - 1 & = & 0.5 \end{array}$$

Therefore, $Z_{3,3} = \{\begin{pmatrix}0.8\\0\\0\end{pmatrix}\}$.

In this case, there are two minimal covering of the set $Z_3 = \{Z_{1,3}, Z_{2,3}, Z_{3,3}\}$*:*

$$\begin{pmatrix}0.8\\0.4\\0\end{pmatrix} = \begin{pmatrix}0.8\\0\\0\end{pmatrix} \vee \begin{pmatrix}0\\0.4\\0\end{pmatrix} \text{ and } \begin{pmatrix}0.8\\0\\0.6\end{pmatrix} = \begin{pmatrix}0.8\\0\\0\end{pmatrix} \vee \begin{pmatrix}0\\0\\0.6\end{pmatrix}$$

This yields four fuzzy relations, defined as follows

X_4	w_1	w_2	w_3
v_1	0	0	0.8
v_2	0	0.8	0.4
v_3	0.9	0.5	0

X_5	w_1	w_2	w_3
v_1	0	0	0.8
v_2	0	0.8	0
v_3	0.9	0.5	0.6

X_6	w_1	w_2	w_3
v_1	0	0.5	0.8
v_2	0	0.8	0.4
v_3	0.9	0	0

X_7	w_1	w_2	w_3
v_1	0	0.5	0.8
v_2	0	0.8	0
v_3	0.9	0	0.6

that solve Equation (2). By their construction and the properties of the Łukasiewicz conjunctor, they are minimal solutions.

Example III.2. *In this example, we consider the Gödel structure [9], then* $L = [0, 1]$ *and* $\odot\colon L \times L \to L$ *and* $\to\colon L \times L \to L$ *are defined by* $x \odot y = \min\{x, y\}$ *and*

$$y \to z = \begin{cases} 1 & \text{if } y \leq z \\ z & \text{otherwise} \end{cases}$$

for all $x, y, z \in [0, 1]$*. Given* $U = \{u_1, u_2\}$*,* $V = \{v_1, v_2, v_3\}$*,* $W = \{w\}$ *and*

R	v_1	v_2	v_3
u_1	0.6	0.4	0.5
u_2	0.8	0.7	0.6
u_3	0.9	1	0.9

T	w
u_1	0.6
u_2	0.7
u_3	0.9

the direct computation shows that

$R \Rightarrow T$	w
v_1	0.7
v_2	0.9
v_3	1.0

is the maximal solution of Equation (2). In order to verify the equality $(R \circ (R \Rightarrow T))\langle u_1, w\rangle = 0.6$ *we compute*

$$\begin{array}{lcccl} R\langle u_1, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w\rangle & = & 0.6 \wedge 0.7 & = & 0.6 \\ R\langle u_1, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w\rangle & = & 0.4 \wedge 0.9 & = & 0.4 \\ R\langle u_1, v_3\rangle \odot (R \Rightarrow T)\langle v_2, w\rangle & = & 0.5 \wedge 1.0 & = & 0.5 \end{array}$$

Note that we only need the value associated with v_1*. Moreover, this value can be reduced until* 0.6*. Hence, the first (and only) column of a solution has to be contained in the following set*

$$Z_{1,1} = \{\begin{pmatrix}x\\0\\0\end{pmatrix} \mid 0.6 \leq x \leq 1\}$$

Focusing on our main goal, the least one is the column associated with a minimal solution. Hence, we only consider the column $Z_{1,1} = \{\begin{pmatrix}0.6\\0\\0\end{pmatrix}\}$

For $(R \circ (R \Rightarrow T))\langle u_2, w\rangle = 0.7$ *we have*

$$\begin{array}{lcccl} R\langle u_2, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w\rangle & = & 0.8 \wedge 0.7 & = & 0.7 \\ R\langle u_2, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w\rangle & = & 0.7 \wedge 0.9 & = & 0.7 \\ R\langle u_2, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w\rangle & = & 0.6 \wedge 1.0 & = & 0.6 \end{array}$$

Now the values associated with v_1 *and* v_2 *provide the maximum. Furthermore, the value for* v_2 *can also be decreased, specifically, any element* x *in* $[0.7, 1]$ *provides the same maximum result:* $0.7 \wedge x = 0.7$*. Therefore, focusing on the minimal solutions we only need to consider:*

$$Z_{2,1} = \{\begin{pmatrix}0.7\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0.7\\0\end{pmatrix}\}$$

Finally, for $(R \circ (R \Rightarrow T))\langle u_3, w\rangle = 0.9$ *we calculate*

$$\begin{array}{lcccl} R\langle u_3, v_1\rangle \odot (R \Rightarrow T)\langle v_1, w\rangle & = & 0.9 \wedge 0.7 & = & 0.7 \\ R\langle u_3, v_2\rangle \odot (R \Rightarrow T)\langle v_2, w\rangle & = & 1.0 \wedge 0.9 & = & 0.9 \\ R\langle u_3, v_3\rangle \odot (R \Rightarrow T)\langle v_3, w\rangle & = & 0.9 \wedge 1.0 & = & 0.9 \end{array}$$

In this last case v_2 *and* v_3 *are involved in the computation of the maximum and the value associated with* v_3 *can be decreased until* 0.9*. These considerations yield the following minimal solutions*

$$\begin{pmatrix}0.6\\0.7\\0.9\end{pmatrix}, \begin{pmatrix}0.7\\0\\0.9\end{pmatrix}, \begin{pmatrix}0.6\\0.9\\0\end{pmatrix}$$

IV. Conclusion and future works

The main aim of this research is to define as generally as possible an algebraic structure that allows the existence of minimal solutions of the fuzzy relation equations defined based on this structure. For that, a general increasing operation $\odot$, which only satisfies the adjointness property, i.e. is residuated, and satisfies the IPNE-condition, has been considered to define a general fuzzy relation equation, which has minimal solutions whenever a solution exists. Moreover, a new algebraic characterization using the notion of covering is introduced,

which provides a method to obtain the minimal solutions and, consequently, the whole set of solutions.

As future work, the obtained results will be applied to several problems in fuzzy logic, such as to abduction reasoning. It is well-known that implications in MV-algebras are infinitely distributive. A topic of future study is to characterize all structures where implication is infinitely distributivity. Algebraic structures that satisfy the INPE-condition are not studied much; also they will be a topic of future research.

REFERENCES

[1] W. Bandler and L. Kohout. Semantics of implication operators and fuzzy relational products. *Int. J. Man-Machine Studies*, 12:89–116, 1980.

[2] E. Bartl. Minimal solutions of generalized fuzzy relational equations: Probabilistic algorithm based on greedy approach. *Fuzzy Sets and Systems*, 260(0):25 – 42, 2015.

[3] R. Bělohlávek. *Fuzzy Relational Systems: Foundations and Principles*. Kluwer Academic Publishers, 2002.

[4] L. Chen and P. Wang. Fuzzy relation equations (ii): The branch-point-solutions and the categorized minimal solutions. *Soft Computing–A Fusion of Foundations, Methodologies and Applications*, 11:33–40, 2007.

[5] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. Multi-adjoint algebras versus non-commutative residuated structures. *International Journal of Approximate Reasoning*, 66:119–138, 2015.

[6] B. De Baets. Analytical solution methods for fuzzy relation equations. In D. Dubois and H. Prade, editors, *The Handbooks of Fuzzy Sets Series*, volume 1, pages 291–340. Kluwer, Dordrecht, 1999.

[7] J. C. Díaz-Moreno, J. Medina, and M. Ojeda-Aciego. On basic conditions to generate multi-adjoint concept lattices via galois connections. *International Journal of General Systems*, 43(2):149–161, 2014.

[8] J. C. Díaz-Moreno, J. Medina, and E. Turunen. Minimal solutions of general fuzzy relation equations on linear carriers. an algebraic characterization. *Fuzzy Sets and Systems*, pages –, 2016.

[9] P. Hájek. *Metamathematics of Fuzzy Logic*. Trends in Logic. Kluwer Academic, 1998.

[10] J.-L. Lin, Y.-K. Wu, and S.-M. Guu. On fuzzy relational equations and the covering problem. *Information Sciences*, 181(14):2951–2963, 2011.

[11] E. Sanchez. Resolution of composite fuzzy relation equations. *Information and Control*, 30(1):38–48, 1976.

[12] B.-S. Shieh. Solution to the covering problem. *Information Sciences*, 222(0):626–633, 2013.

[13] E. Shivanian. An algorithm for finding solutions of fuzzy relation equations with max-lukasiewicz composition. *Mathware and Soft Computing*, 17:15–26, 2010.

[14] E. Turunen. On generalized fuzzy relation equations: necessary and sufficient conditions for the existence of solutions. *Acta Universitatis Carolinae. Mathematica et Physica*, 028(1):33–37, 1987.

[15] Q.-Q. Xiong and X.-P. Wang. Solution sets of fuzzy relational equations on complete Brouwerian lattices. *Information Sciences*, 177(21):4757–4767, 2007.

[16] Q.-Q. Xiong and X.-P. Wang. Fuzzy relational equations on complete Brouwerian lattices. *Information Sciences*, 193(0):141–152, 2012.

[17] C.-T. Yeh. On the minimal solutions of max-min fuzzy relational equations. *Fuzzy Sets and Systems*, 159(1):23–39, 2008.

[18] Z. Zahariev. `http://www.mathworks.com/matlabcentral, fuzzy-calculus-core-fc2ore`, 2010.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 25–32
DOI: 10.15439/2016F372
ACSIS, Vol. 9. ISSN 2300-5963

Identification of Product's Features Based on Customer Reviews

Katarzyna Smietanka
University of Warsaw
Banacha 2, 02-097 Warszawa, Poland
katarzyna.smietanka@students.mimuw.edu.pl

***Abstract*—In recent years an e-commerce has become more and more popular. This fact is mainly related to a low cost of running a business, wide access to a large group of potential customers and ease of advertising. Analysis of products' reviews can lead to valuable insights for both customers and manufacturers. Owing to positive reviews a future customer may be convinced to buy the product. A number of reviews for one product can amount to even hundreds what makes it hard for a potential buyer to read them all. The main aim of this paper is to present a method for mining reviews considering products' features, extracting products' features and preparing a summary of reviews. For that purpose a new promising technique—*Rule-Based Similarity Model* is used. The performance of the algorithm has been verified on online product review articles.**

***Keywords*—opinion mining; data mining; Rule-Based Similarity Model; Natural Language Processing; sentiment mining;**

I. Introduction

THE utilization of websites aggregating customer's reviews as a source of information has increased rapidly with the expansion of e-commerce. More and more people decide to buy, even everyday products on the Internet. In order to enhance customer shopping experience and satisfaction, merchants give a possibility to share feedback. Usually, such a form of a review is unlimited and unstructured which encourages customers to express opinion of a product or a service. There are lots of websites on which we can find customers' opinions about products or services. The most popular are *Amazon*, *Google Shopping*, *Google Maps* - reviews about places and *OPiNEO* - a polish web page aggregating products' reviews. As more and more people are eager to write reviews, some popular products or services can get a huge number of reviews. Some of these written reviews can be extremely long, but only a few sentences focus on product's features. Taking into account the unlimited form of a feedback and a huge number of reviews, it may be difficult for a potential customer to analyse all of them. It justifies the need for an automatic system to extract key sentences from reviews considering products' features.

The knowledge about others people's opinion is not only beneficial for a potential buyer, but it is also useful for manufacturers. They can keep track of both positive and negative aspects of a product. In a case of negative opinions, manufacturers can decide to improve the quality of the product to enhance people's satisfaction. Thanks to this, it is possible to decrease the number of future complaints. Moreover, it helps to build a positive image of the products and services.

Usually feature extraction algorithms in customers' reviews are decomposed into three subtasks: (i) identification of features of a product, (ii) identification of opinions expressed about features of a product (iii) determination a sentiment orientation of the opinions [1].

II. Related Works

In [2] K. Kahn, B. Baharudin and A. Khan proposed hybrid patterns technique for features identification. The hybrid pattern is a combination of dependency patterns, which exploits grammatical structure and contextual rules. Most of the patterns derive from observation and empirical analysis. Patterns: *NN, NN NN, JJ NN, NN NN NN, JJ NN NN, JJ JJ NN* (*NN* - noun, *JJ* - adjective) are potentially exploited for candidate selection of product feature. Additionally authors created new groups of patterns: *Definite Base Noun* (definite article *the* before the noun phrase), *Linking Verb Based Noun Phrases* (based on assumption that linking verbs between *JJ* and *NN* provide best clues for opinion expressions) and *Preposition Based Noun Phrases* (noun phrases with *of* preposition).

Hiu and Liu in [3] presented a system which uses association rule mining [4] to extract nouns and noun phrases as feature candidates from product reviews. The main parts of their opinion summary system are: (i) part-of-speech tagging, (ii) frequent feature generation (iii) feature pruning (iv) opinion word extraction (v) infrequent feature identification [1]. In (i) stage, linguistic parser identifies noun and noun phrases. For generation frequent features (ii) authors use found expressions in previous stage. Other components of a sentence than noun/noun phrases are unlikely to be product features. In order to find frequent features, authors run [5] that bases on *association rule mining* technique. The main purpose of (iii) is to remove uninteresting and redundant features generated by [5] algorithm. There are two types of pruning: *compactness pruning* - remove those candidate features whose words do not appear together and *redundancy pruning* - remove redundant features that contain single word (e.g. *life* is not a useful feature while *battery life* is meaningful feature). In (v) is used an idea described in Section IV (2). The last stage of the system is identification of opinion words in sentences and semantic orientation using *bootstrapping* technique and *WordNet*.

In paper [6], authors presented aspect-based summary model, where a summary is built by extracting relevant aspects of a service. Partially that approach bases on ideas from [3]. A novel aspect of created models is that they exploit user-provided labels and domain-specific characteristics of service reviews to increase quality.

Authors A. Ghobadi and M. Rahgozar [7] approached the problem using an ontology to extract the products' information. Ontology is defined as concepts, their relationships and concepts instances of a specific domain. Required information is automatically extracted from dependency patterns.

In [8] authors proposed a method for aspect extraction based on *Double Propagation* with aspect recommendations: semantic similarity-based and aspect association-based. Semantic similarity-based recommendation aims to solve the problem of missing synonymous aspects of DP using word vectors from a large corpus. Aspect association-based recommendation is based on an idea that many aspects are correlated or co-occur across domains.

III. Analysis of the problem

A. Difficulties in analysing product reviews

The main problem connected with identifying sentences containing product's features is the unstructured form of reviews. Usually, reviews are rather long, but they only partially cover the topics regarding real aspects of a service or a product. Moreover, most of the opinions are written in an informal way, it often happens that it is not consistent with language grammar. Authors of reviews use different types of abbreviations (*g8* - meaning great, or *imo* - in my opinion) and make a digression, what significantly complicates a task of researching reviews. Another problem is connected with mapping implicit aspect expressions to aspects of a product. It is very hard to properly extract and identify that aspect, e.g. *fit in pockets* in the sentence *"This phone will not easily fit in pockets."* [9].

The problem of aspect and entity coreference resolution is also present in the task of feature identification. It has been extensively researched recent years. A coreference determines which mentioned entities or aspects refer to the same entity. For example in sentences: *"The Sony camera is better than the Canon camera. It is cheap too."*, *it* refers to *Sony*, because of the way how people express their opinions [9].

B. Review formats

There are several types of review formats - collecting customers' opinions about the product. The most popular are: (a) *pros and cons* - a user separately describes positive and negative sides of the product, (b) *pros, cons and review* - despite of pros and cons opinion, user is asked to write whole review (c) *unlimited format* - only text of the review without explicit division into positive and negative aspects (d) *questionnaires* - a structured opinion, the user expresses his opinion by while choosing one of available option. As it is suggested in [10], the review format should be taken into consideration choosing a proper algorithm for extracting features. In (a) and (c) reviewers usually use full sentences, but section *pros and cons* of (b) contains concise and short expressions. The precise and accurate format of (d) makes it easy to analyse and aggregate users' opinions.

C. Main types of opinions

In [11], *Jindal* and *Liu* proposed two types of opinions: (i) a regular opinion, (ii) a comparative opinion . In (i) sentiment expressions refer to some target entities. For example the sentence *"The touch screen is really cool"* presents the direct regular opinion about *the touch screen*, whereas the sentence *"After taking the drug, my pain has gone"* is an example of an indirect regular opinion. In comparative opinion (ii) a part of product is commented in relation to other one, e.g. *"iPhone is better than Blackberry"*.

D. Assumptions

This research is mainly focused on the problem of mining reviews containing features of a product, extracting those features and preparing a structured summary based on customers' reviews. The summary looks like the following:

```
Product Name: DVD_player
  1. Feature: size
    Number of extracted sentences: ...
    Individual reviews: ...
  2. Feature: ...
    Number of extracted sentences: ...
    Individual reviews: ...
```

where `Individual reviews` points to the specific sentences from reviews that comment about the feature. We are interested in all features of the product mentioned by customers, independently of the opinion polarity. In this paper, only unlimited format of reviews is considered and the focus is put on regular opinions.

IV. Aspect Extraction - Approaches

The problem of identification of products' features is strictly connected with aspect extraction, which is a part of an information extraction task. Focusing on the explicit aspect extraction, in [11] authors proposed four main approaches:

1) Extraction based on frequent nouns and noun phrases.
2) Extraction by exploiting opinion and target relations.
3) Extraction using supervised learning.
4) Extraction using topic modeling.

The (1) approach finds aspects that are nouns and noun phrases from a large number of reviews in a given domain. Nouns and noun phrases are identified by a part-of-speech tagger and their occurrence frequencies are counted. The way of solving the problem is justified by the observation that people commenting on products use similar vocabulary. Moreover, it is based on the assumption that frequently talked nouns are important and genuine features of the product. In (2), the relationships between targets of opinion are exploited to extract aspects. It is based on an idea that the same sentiment word can be used to describe or modify different aspects [3]. Supervised learning algorithms (3) need manually labeled data for training. The most dominant methods are *Hidden Markov Models* and *Conditional Random Fields*. Additionally, supervised learning methods usually use a set of domain-independent characteristics such as word distance, syntactic dependency, tokens or part-of-speech tagging. (4) is an unsupervised learning method, which uses statistical topic models. It assumes that each document consists of a mixture of topics, each topic is a probability distribution over words. Two main models are *Probabilistic Latent Semantic Analysis* and *Latent Dirichlet Allocation*.

V. Introduction to Rough Sets

Rule-Based Similarity Model that is presented in Section VI derives from the theory of rough sets (definitions presented in this section comes from [12]). The theory of rough sets, proposed by Zdzislaw Pawlak in 1981 [13], provides a mathematical formalism for reasoning about imperfect data and knowledge. In the rough set theory, available knowledge about object $u \in U$ (U is a finite non-empty set of objects) is represented as a vector of information about values of its *attributes*. An attribute can be treated as a function $a : U \rightarrow V_a$ that assigns values from a set V_a to objects from U. All available information about objects from U can be stored in a structure called an *information system* - $\mathbb{S}$, which is defined as a tuple $\mathbb{S} = (U, A)$, where A is a finite non-empty set of attributes.

Conditional attributes are those attributes about which information about values of all attributes from A can be obtained for any object, including those which are not present in U. A *decision attribute* is an attribute, which can be used to define a partitioning of U into disjoint sets. An information system with a defined decision attributes is called a *decision system* and is denoted by $\mathbb{S}_d = (U, A \cup \{d\})$, where $A \cap \{d\} = \varnothing$.

The rough set theory is often utilized to provide description of concepts from the considered universe. Any concept can generally be associated with a subset of objects from U which belong or match to it. Decision attributes in a decision system can usually be interpreted as expressing the property of belongingness to some concept. Given some information about characteristics (values of attributes) of objects corresponding to considered concept, it can be described using a *decision logic language*. Decision logic language L_A is defined over an alphabet consisting of a set of attribute constants (names of attributes from A) and a set of attribute value constants (symbols representing possible attribute values). Atomic formulas of L_A are attribute-value pairs $a = v$, where $a \in A$ and $v \in V_a$. Each description (a formula) ϕ in a decision logic language L_A can be associated with a set of objects from U that satisfy it.

Knowledge about dependencies between conditional attributes and decision attributes of a decision system are often represented using special formulas called *decision rules*.

Definition V.1 (Decision rules)**.** *Let A and D be conditional and decision attribute sets of some decision system and let $L_{A\cup D}$ be a decision logic language and π be a formula of $L_{A\cup D}$. We will say that π is a decision rule iff the following conditions are met: (i) $\pi = (\phi \rightarrow \psi)$, (ii) ϕ and ψ are conjunctions of descriptors, (iii) ϕ is a formula of L_A and ψ is a formula of L_D.*

The right hand side of a decision rule $\pi = (\phi \rightarrow \psi)$ (i.e. ψ) is called a *successor* of a rule (denoted by $rh(\pi)$) and the left hand side (denoted by $lh(\pi)$) is called a *predecessor* (i.e. ϕ).

There is also a different type of rules - *inhibitory rules*, which is useful for analysing dependencies in data with multiple decision values.

Definition V.2 (Inhibitory rules)**.** *Let A and D be conditional and decision attribute sets of a decision system and let $L_{A\cup D}$ be a decision logic language and π be a formula of $L_{A\cup D}$. We will say that π is an inhibitory rule iff the following conditions are met: (i) $\pi = (\phi \rightarrow \neg\psi)$, (ii) ϕ and ψ are conjunctions of descriptors, (iii) ϕ is a formula of L_A and ψ is a formula of L_D.*

An inhibitory rule tells us that an object which satisfies the predecessor of this rule cannot belong to a pointed decision class. The inhibitory rules can be seen as a complement to decision rules as they often provide means to classify objects which are difficult to cover by the traditional rules.

Usefulness of a rule for prediction of decision classes of new objects can be quantitatively assessed using rule quality measures: *support* and *confidence*. The support of a rule π is defined as

$$supp(\pi) = \frac{|lh(\pi)(U)|}{|U|} \tag{1}$$

and the confidence of π is:

$$conf(\pi) = \frac{|lh(\pi)(U) \cap rh(\pi)(U)|}{|lh(\pi)(U)|}. \tag{2}$$

In the rough set theory any arbitrary set of objects X can be approximated within an information system $\mathbb{S} = (U, A)$ by a pair of definable sets $App(X) = (\underline{X}, \overline{X})$, called a rough set of X in S. The set $\underline{X}$ is the largest and the set $\overline{X}$ is the smallest definable set which contains X. The sets $\underline{X}$ and $\overline{X}$ are called a *lower* and *upper approximation* of X in S, respectively.

VI. Rule-Based Similarity Model

In [12], author proposed a similarity model, called Rule-Based Similarity (RBS). In RBS the similarity is assessed by examining whether two objects share some binary higher-level features. Features which are relevant for a considered similarity context are extracted from data and their importance is assessed based on available data.

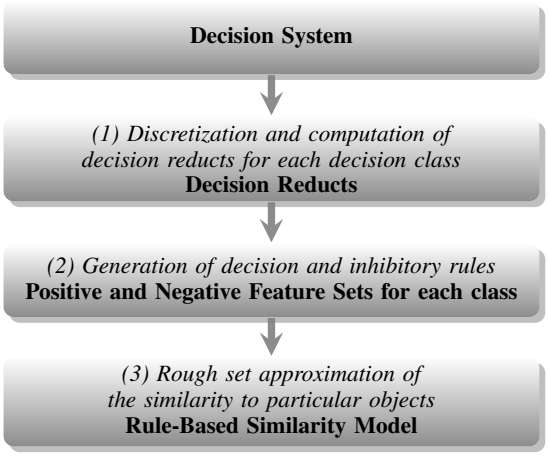


Fig. 1. A construction schema of the RBS model [12].

In the presented RBS model similarity is learnt in a classification context, which is defined by a decision attribute in a data set. In the (1) step of the model, raw attributes values are transformed into a symbolic representation that resembles

basic qualitative characteristics of objects. This approach is based on a way in which people assess similarity. Tversky [14] suggested that people rarely think in terms of exact numbers but rather tend to operate on binary characteristics (e.g. *an object is large* or *an object is round*). In a practical data analysis, it is enough to divide numerical attribute values into intervals representing qualitative symbols. This process is done by applying a heuristic discretization technique. For each decision class are computed *decision reducts*. Reducts are an important tool for selecting informative features in the rough set theory, they help to discover redundancy and dependencies between attributes. A *decision reduct* is a subset of attributes that discriminates objects from different decision classes and is minimal in a sense that no attribute can be removed from a reduct without losing the properties of discernibility.

In (2), higher-level features that are important for a notion of similarity are created using a rule mining algorithm. Each of those features is defined by the left-hand side of a rule - $lh(\pi)$. There are two types of rules: *decision rules* and *inhibitory rules*, described in Section V. Decision rules aim is to provide partial descriptions of concepts indicated by the decision attribute. They are used to predict decision classes of new objects. Inhibitory rules specify what kind of objects cannot belong to a pointed decision class. Depending on a type of a rule, the corresponding feature can be useful either as an argument for or against the similarity to a matching object.

The induction of rules in RBS may be treated as a process of learning aggregations of local similarities from data. Features defined by predecessor of the rules express higher-level properties of objects.

The last stage includes rough set approximation of the similarity to particular objects. The problem of learning the similarity relation in RBS is closely related to searching for a relevant approximation space. Formally, let $F^{+}_{(i)}$ and $F^{-}_{(i)}$ be the sets of binary features derived from the decision and the inhibitory rules, generated for i-th decision class:

$$F^{+}_{(i)} = \{\phi : (\phi \rightarrow (d = i)) \in RuleSet_i\} \tag{3}$$

$$F^{-}_{(i)} = \{\phi : (\phi \rightarrow \neg(d = i)) \in RuleSet_i\} \tag{4}$$

where $RuleSet_i$ is a set derived from a reduct associated with the i-th decision class. The rule set may be generated using any rule mining algorithm. For efficiency in practical applications of the model it is necessary to require that the generated sets of rules $RuleSet_i$ be minimal. It means that there is no rule $\pi \in RuleSet_i$ that could be removed without reducing the set of covered objects.

A feature ϕ is a decision logic formula, we will say that an object u, described in a decision system $\mathbb{S} = (\mathbb{U}, \mathbb{A})$, has a feature ϕ iff $u \models \phi$. A set of all object from U that have the feature ϕ will be denoted by $\phi(U)$. In RBS a *similarity relation* is approximated by means of approximating multiple concepts of being similar to a specific object. It consists of those objects from U which share with u at least one feature from the set F^{+}_{i}, where i is a decision class of u ($d(u) = i$):

$$SIM_i(u) = \bigcup_{\phi \in F^{+}_{i} \wedge u \models \phi} \phi(U) \tag{5}$$

the approximation of the dissimilarity to u is a set of objects from U which have at least one feature from $F^{-}_{(i)}$ that is *not in common* with u:

$$DIS^{0}_{i}(u) = \bigcup_{\phi \in F^{-}_{i} \wedge u \not\models \phi} \phi(U) \tag{6}$$

The set of objects that have at least one feature from $F^{-}_{(i)}$ that *in common* with u will be denoted by

$$DIS^{1}_{i}(u) = \bigcup_{\phi \in F^{-}_{i} \wedge u \models \phi} \phi(U) \tag{7}$$

The functions SIM and DIS are used for the approximation of the similarity and the dissimilarity to objects from U. The assessment of degree in which an object u_1 is similar and dissimilar to u_2 is done using two functions (abbreviations $SIM(u) = SIM_{(d(u))}(u)$; $DIS(u) = DIS^{0}_{(d(u))}(u)$ are written when the decision for an object u is known):

$$Similarity(u_1, u_2) = \frac{|SIM(u_1) \cap SIM_{d(u_1)}(u_2)|}{|SIM(u_1)| + C_{sim}} \tag{8}$$

$$Dissimilarity(u_1, u_2) = \frac{|DIS(u_1) \cap DIS^{1}_{d(u_1)}(u_2)|}{|DIS(u_1)| + C_{dis}} \tag{9}$$

In the above formulas C_{sim} and C_{dis} are positive constants which can be treated as parameters of the model. The similarity function of the RBS model combines values of $Similarity$ (Sim) and $Dissimilarity$ (Dis) for a given pair of objects:

$$Sim_{RBS}(u_1, u_2) = F(Sim(u_1, u_2), Dis(u_1, u_2)) \tag{10}$$

where $F : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ can be any function that is monotonically increasing with regard to its first argument and monotonically decreasing with regard to its second argument. Detailed description of the similarity function is provided in [12].

VII. Proposed Algorithm

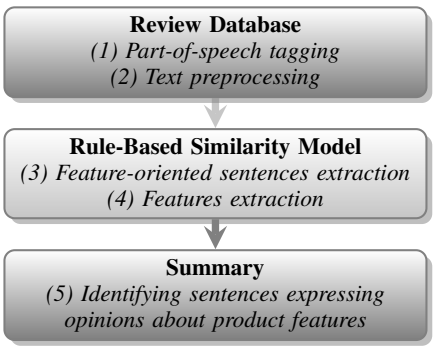


Fig. 2. A construction schema of extraction feature summary system.

Figure 2 provides a general overview of the construction of a summary system for the feature extraction. Product names and reviews associated with a product are an input of the system. Reviews for a single product are split into sentences and tagged with decision class - *positive* if a sentence contains feature otherwise *negative*. This represents the true reference classification. As an output, for each product system creates

a summary of the reviews in the following format: product feature and list of sentences from reviews correlated with that feature.

In (2) preprocessing of reviews is performed, what plays an important role in text mining discipline. Text preprocessing stage includes stop words and punctuation removal, stemming. The feature extraction task is focused on finding features that appear explicitly as noun or noun phrases in the reviews. To enhance the quality of feature-oriented sentences extraction, adjectives and adverbs are extracted. It derives from observation, that usually when people express their opinion about an aspect of a product, they use opinion words such as adjectives and adverbs. To identify noun/noun phrases, adjectives and adverbs (1) from the reviews, part-of-speech tagger - NLTK parser (*Natural Language Toolkit*) is used. For each sentence, identified and preprocessed parts of speech are saved in the review database. The following shows a sentence *"I am very pleased with its quality"* with the POS tags. For instance, a tag NN indicates a noun and JJ indicates an adjective.

$$[(I, NN), (am, VBP), (very, RB), (pleased, JJ), (with, IN), (its, PRP), (quality, NN)] \quad (11)$$

In next step of the process, Rule-Based Similarity Model, which is described in Section VI, is built. Documents, sentences from reviews, are transformed into a document-term matrix which represents the frequency of terms occurring in a collection of documents. This is an input to RBS model. In a document-term matrix, rows correspond to sentences in the collection and columns correspond to terms. To determine the value that each entry in the matrix, we use a *term-frequency* scheme, defined as: $f_{t,d}$ of term t in document (sentence) d is the number of times that t occurs in d. The similarity function of RBS model is used to perform a classification of previously unseen objects of textual documents into groups: *positive* - sentences containing product features otherwise *negative*.

The attributes from the left hand side of *decision rules* - $lh(\pi)$ ($\pi = (\phi \rightarrow \psi)$) calculated in RBS model, represents a set of candidates for product features - arguments for the similarity to a matching objects. The formula ϕ from $lh(\pi)$ is expressed by a compound formula - the attribute value pairs. The value from atomic formula (attribute-value pair) is an interval that represents a required frequency of an attribute (term).

The taken approach is justified by the reviews representation as an input to RBS model, association mining algorithm for creating rules and empirical observation of human expressions. On the other hand, *inhibitory rules* are useful as an argument against the similarity to matching objects. In (5) algorithm identifies sentences from reviews expressing opinions about product features and creates a summary.

The proposed algorithm partly depends on opinion identification - sentiment words (adjectives and adverbs) are an input to RBS model. Moreover, the algorithm uses empirically defined threshold for a minimum support and a maximum length of rules, which are created in an association mining process.

VIII. Experiments

The approach was tested on dataset of customer reviews for five products collected from *Amazon.com* and *Cnet.com* as described in [3]. Dataset contains customer reviews focusing on electronic products: a DVD player - *Apex AD2600 DVD player*, digital cameras - *Canon G3* and *Nikon coolpix*, a cell phone - *Nokia 6610* and a MP3 player - *Creative Labs*.

TABLE I. Product review dataset.

Product Name	No. of reviews	No. of sentences
Digital camera1	45	597
Digital camera2	34	346
Cell phone	41	546
MP3 player	95	1716
DVD player	99	739

Originally dataset was annotated by *Hu* and *Liu* [3], they define a product feature as a characteristic of the product which customers have expressed an opinion about. Opinion is a statement which explicitly defines an attitude towards feature, which is positive or negative. The dataset was reannotated by authors of [1], in order to put the main focus on feature extraction. Annotated terms as features satisfy one of the criteria: (i) part of relationship between product and a feature (ii) attribute of product, e.g. *design* - attribute of camera (iii) attribute of a known feature, e.g. *battery life* - it is an attribute of *battery*. Experiments were conducted on dataset tagged by [1].

TABLE II. Detailed summary of dataset.

Product Name	Dataset tagged by [3]		Dataset tagged by [1]	
	Distinct	Total	Distinct	Total
Digital camera1	100	257	161	594
Digital camera2	74	185	120	340
Cell phone	109	310	140	470
MP3 player	180	736	231	1031
DVD player	110	347	166	519

The effectiveness of proposed algorithm is evaluated by standard evaluation methods: *precision*, *recall* and *f1-score*. Precision answers the question *"How many selected items are relevant?"*, while recall expresses an idea about *"how many relevant items are selected?"*. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. F1-Score is a measure of a test's accuracy, it can be interpreted as a weighted average of the precision and recall. This measure reaches its best value at 1 and its worsts at 0.

TABLE III. Summary of rule sets from RBS model for decision classes.

Product Name	No. of decision rules for positive class	No. of decision rules for negative class
Digital camera1	44	5
Digital camera2	32	9
Cell phone	31	13
MP3 player	98	27
DVD player	38	1
Average	48	11

In Table III is presented a summary of unique rules sets which were generated by the RBS model. It is easy to notice that a number of decision rules for a positive class are several

times greater than a number of decision rules for a negative class. It is mainly connected with empirically set thresholds for a minimum negative and positive confidence for association mining rules. In order to increase a recall measure, minimum confidence for the positive is smaller than for the negative decision class.

A. The effectiveness of feature-oriented sentence extraction

In the task of mining reviews containing product's features (*feature-oriented sentence extraction*), received results from proposed algorithm (RBS similarity from Section VII) are compared with *Support Vector Machine* (SVM) [15] *with hidden concepts* (SVM with concepts). The idea of SVM algorithm is to find boundaries in the input feature space. In order to obtain good performance of SVM in the classification process, before training a classifier, dimension reduction is performed and documents are represented in a space of higher-level terms. For that purpose is used the typical approach as Singular Value Decomposition (SVD), which goal is to find a representation of document term matrix as a product of lower-rank matrices. Using calculated SVD matrix and selected hidden concepts, which are two-terms phrases, documents are represented in a space of that concepts. The name of that technique is *Latent semantic analysis* (LSA) [16], it is a method for extracting and representing contextual-usage of words by statistical computations applied to a large corpus of text. In the following experiment, SVM algorithm is applied using linear kernel. In [17] author justified a good performance of SVM in a text classification context. Combining LSA with SVM is a common method for text classification [17].

As in [12], the quality of the compared models was assessed using two different measures - *mean accuracy* (Mean) and *balanced accuracy* (Balanced). The mean classification accuracy, defined as:

$$Mean = \frac{|\{t \in DataSet : p(t) = d(t)\}|}{|DataSet|} \tag{12}$$

where $DataSet$ is a set of test objects; $p(t)$ is a predication of a decision class for an object t and $d(t)$ is an expected decision class for an object t, was estimated using 3-fold cross-validation technique [18]. The balanced accuracy is calculated by computing standard classification accuracies $Mean_i$ for each decision class and then averaging the result over classes: $i \in \{positive,\ negative\}$ [12].

$$Mean_i = \frac{|\{t \in DataSet : p(t) = d(t) = i\}|}{|\{t \in DataSet : d(t) = i\}|} \tag{13}$$

$$Balanced = \frac{Mean_{positive} + Mean_{negative}}{2} \tag{14}$$

TABLE IV. COMPARISON OF MEAN AND BALANCED ACCURACY FOR FEATURE-ORIENTED SENTENCE EXTRACTION.

Product Name	Proposed algorithm from Section VII		SVM with hidden concepts	
	Mean	Balanced	Mean	Balanced
Digital camera1	0.724	0.704	0.427	0.494
Digital camera2	0.699	0.675	0.396	0.517
Cell phone	0.679	0.685	0.460	0.488
MP3 player	0.715	0.714	0.536	0.511
DVD player	0.743	0.744	0.539	0.526
Average	0.712	0.704	0.472	0.507

For all the product reviews in the classification problem, RBS model achieved better results than SVM with concepts. Both an average *mean accuracy* and *balanced accuracy* measures are higher for RBS model. The statistical significance of mean differences in results between classification algorithms was verified using *a paired t-test*. A null hypothesis was tested that obtained mean measurements for data sets have equal means. Difference in means is considered significant if *p-value* of the test is lower than 0.01. For presented results, the null hypothesis is rejected, the performance of RBS similarity model is better than for SVM model with hidden concepts.

B. The effectiveness of feature extraction

In order to illustrate the effectiveness of proposed feature extraction technique, gained results are compared with widely available Content Term Extraction (CTE) algorithm [19] (similarly as in [9]). As proposed algorithm disregards the original ordering of the terms in sentences, it is impossible to directly evaluate gained results. Evaluation of extracted features is performed on single-word features, multi-word features from annotation are divided into separate terms.

The proposed approach to problem from Section VII is assessed by two different measures: *global quality criteria* and *feature-oriented quality criteria*. The global quality criteria examine the algorithm's performance on the task of extracting features from the collection of reviews. This relates to the main task of creating the summary of features, which involves: identifying features of the product that customers expressed their opinion on and finding review sentences corresponding to extracted features. The feature-oriented quality criteria are focused on an evaluation of how many unique features defined by an expert was found by the algorithm. For that measure, it is not important from which sentence and document an algorithm extracted a certain feature.

TABLE V. COMPARISON OF RECALL AND PRECISION FOR GLOBAL QUALITY CRITERIA IN DATASET TAGGED BY [1].

Product Name	Proposed algorithm from Section VII		Content Term Extraction	
	Precision	Recall	Precision	Recall
Digital camera1	0.610	0.545	0.240	0.485
Digital camera2	0.586	0.545	0.292	0.586
Cell phone	0.562	0.528	0.277	0.531
MP3 player	0.633	0.631	0.195	0.390
DVD player	0.685	0.644	0.209	0.388
Average	0.615	0.579	0.243	0.476

Table V shows the comparison of recall and precision for *global quality criteria* for two different algorithms: proposed algorithm based on RBS model (Section VII) and Content Term Extraction. It can be observed that both of these measures are lower for CTE algorithm. The major reason of poor precision is the fact that CTE generates a large number of terms, which are not product features.

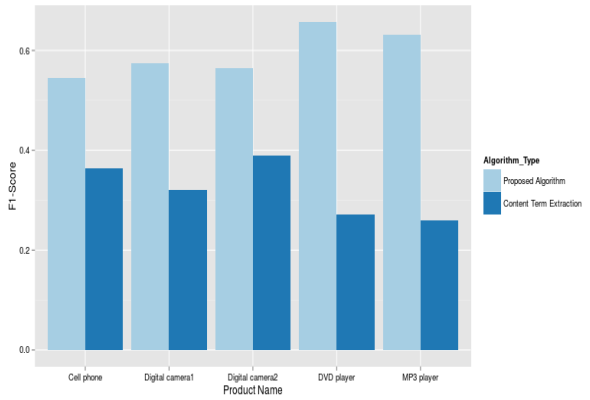


Fig. 3. Comparison of F1-Score for algorithms.

Comparing results with *Likelihood Test approach* presented in [1] (*Table 8. Feature extraction results on instance level*), the proposed algorithm receives better average recall, but the average precision is slightly lower. In contrast to *Association Mining approach* [1], the average precision is improved for the algorithm based on RBS model. The proposed algorithm's performance on the task of identifying features from the collection of reviews turned out to be quite promising.

TABLE VI. COMPARISON OF RECALL AND PRECISION FOR FEATURE-ORIENTED QUALITY CRITERIA IN DATASET TAGGED BY [1].

Product Name	**Proposed algorithm from Section VII**		**Content Term Extraction**	
	Precision	Recall	Precision	Recall
Digital camera1	0.811	0.088	0.135	0.750
Digital camera2	0.806	0.107	0.164	0.805
Cell phone	0.673	0.083	0.149	0.808
MP3 player	0.703	0.110	0.115	0.775
DVD player	0.723	0.112	0.145	0.730
Average	0.743	0.100	0.142	0.774

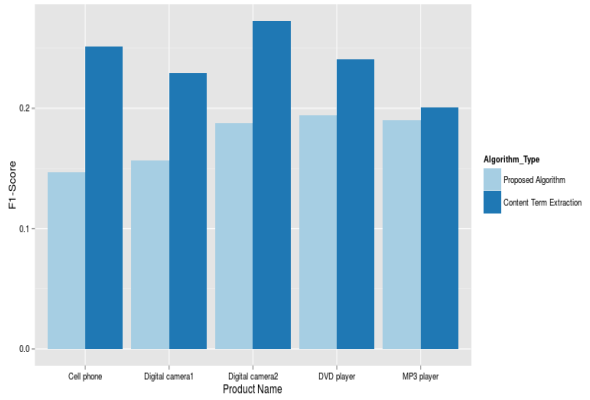


Fig. 4. Comparison of F1-Score for algorithms.

Considering *feature-oriented quality criteria* it can be noticed that the *F1-Score* measure for product reviews is better for CTE algorithm. However, the difference between that measure for the proposed and CTE algorithm is not so meaningful. The low recall for the proposed algorithm is a result of a construction of a decision reduct, which is a minimal subset of attributes that discriminates objects from different decision classes. The high recall for CTE approach derives from the fact that CTE generates a lot of terms, but it results in a low precision. The average precision for the proposed algorithm is about 0.743, whereas for CTE algorithm is 0.142. *Likelihood Test Approach* [1] (LTA) has a similar property - a high precision and a low recall as the algorithm based on RBS. The LTA average recall is about 0.104 and the precision amounts to 0.804.

IX. CONCLUSIONS

In this paper, the problem of identification of product's features based on customer reviews and the new method which based on Rule-Based Similarity Model were discussed. Moreover, difficulties in analysing product reviews, review formats and methods for extracting product features were described. In particular, there were presented results of conducted experiments in comparison to other algorithms and evaluated gained results with two measures: *global quality criteria* and *feature-oriented quality criteria*.

Additionally, this paper shows that algorithm based on Rule-Based Similarity Model could be successfully used for the problem of feature extraction and feature-oriented sentence extraction problem. Proposed algorithm does not require any prior knowledge despite labelled data to train the RBS model.

In the future, this work will be extended with evaluation of the performance of the proposed algorithm on other data sets. The main focus will be put on improving the recall of the described method with regard to the feature-oriented quality criteria.

REFERENCES

[1] L. Ferreira, N. Jakob, and I. Gurevych, "A comparative study of feature extraction algorithms in customer reviews," in *Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA*, 2008, pp. 144–151. [Online]. Available: http://dx.doi.org/10.1109/ICSC.2008.40

[2] K. Khan, B. Baharudin, and A. Khan, "Identifying product features from customer reviews using hybrid patterns," *Int. Arab J. Inf. Technol.*, vol. 11, no. 3, pp. 281–286, 2014. [Online]. Available: http://www.ccis2k.org/iajit/index.php?option=com_content&task=blogcategory&id=92&Itemid=353

[3] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*, 2004, pp. 755–760. [Online]. Available: http://www.aaai.org/Library/AAAI/2004/aaai04-119.php

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, 1994, pp. 487–499. [Online]. Available: http://www.vldb.org/conf/1994/P487.PDF

[5] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," 1998, pp. 80–86.

[6] S. Blair-goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar, "Building a sentiment summarizer for local service reviews," in *In NLP in the Information Explosion Era*, 2008.

[7] A. Ghobadi and M. Rahgozar, "An ontology-based semantic extraction approach for b2c ecommerce."

[8] Q. Liu, B. Liu, Y. Zhang, D. S. Kim, and Z. Gao, "Improving opinion aspect extraction using semantic similarity and aspect associations," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, 2016, pp. 2986–2992. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11973

[9] C. C. Aggarwal and C. X. Zhai, *Mining Text Data*. Springer Publishing Company, Incorporated, 2012.

[10] M. Hu and B. Liu, "Opinion feature extraction using class sequential rules," in *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, 2006, pp. 61–66. [Online]. Available: http://www.aaai.org/Library/Symposia/Spring/2006/ss06-03-013.php

[11] B. Liu, *Sentiment Analysis and Opinion Mining*, ser. Synthesis digital library of engineering and computer science. Morgan & Claypool, 2012. [Online]. Available: https://books.google.pl/books?id=Gt8g72e6MuEC

[12] A. Janusz, "Algorithms for similarity relation learning from high dimensional data," *Trans. Rough Sets*, vol. 17, pp. 174–292, 2014. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-54756-0_7

[13] Z. Pawlak, "Information systems theoretical foundations," *Inf. Syst.*, vol. 6, no. 3, pp. 205–218, 1981. [Online]. Available: http://dx.doi.org/10.1016/0306-4379(81)90023-5

[14] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[15] "Support vector machines," https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf, accessed: 2016-05-04.

[16] T. K. Landauer and S. T. Dumais, "Latent semantic analysis," *Scholarpedia*, vol. 3, no. 11, p. 4356, 2008. [Online]. Available: http://www.scholarpedia.org/article/Latent_semantic_analysis

[17] T. Joachims, "Text categorization with suport vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, ser. ECML '98. London, UK, UK: Springer-Verlag, 1998, pp. 137–142. [Online]. Available: http://dl.acm.org/citation.cfm?id=645326.649721

[18] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. [Online]. Available: http://www.jmlr.org/papers/v7/demsar06a.html

[19] "Content term extraction using pos tagging," https://pypi.python.org/pypi/topia.termextract/, accessed: 2016-05-04.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 33–40
DOI: 10.15439/2016F571
ACSIS, Vol. 9. ISSN 2300-5963

Recognition of Compound Objects Based on Network of Comparators

Łukasz Sosnowski
Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
Dituel Sp. z o.o.
Ostrobramska 101 lok. 206, 04-041 Warsaw, Poland
Email: Lukasz.Sosnowski@ibspan.waw.pl

Marcin Szczuka
Institute of Informatics, University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
Email: szczuka@mimuw.edu.pl

***Abstract*—This paper proposes a methodology for compound objects' recognition based on comparators and comparator networks. The methodology is supported by a collection of techniques and algorithms for construction and learning of comparator networks. Formal description of the methodology is accompanied by selected examples of its application in real-life problems. The described methodology has been implemented as a software library and may be used for a variety of future applications.**

I. Introduction

IN THIS position paper we propose an approach to computational tasks which involve management, identification, classification and modelling of data objects that are intrinsically complex, compound and described by means of various types of information. Our proposed approach is based on *comparators* and *comparator networks*. A comparator is a basic, relatively simple computational element that models local similarity between objects of possibly compound nature, such as phenomena, processes or sub-systems. They can be arranged into comparator networks that are capable of aggregating and summarizing local information about similarity into a global measure of proximity between data objects. This versatility provides the ability to solve the problem at hand by decomposing it into simpler, localised steps that can be processed quickly. Then, local results from elementary units can be aggregated and processed in the network producing the overall, possibly complex and multi-dimensional final result.

The overall purpose of both the comparator and the network is to generate the vector of numerical values that indicates the levels of various similarities between the object provided as an input and the already processed, well-known *reference objects*. The reference set, i.e. a set consisting of reference objects, can be regarded as the underlying knowledge base. For example, if we attempt to make an assessment of the situation on the road and associated risks at the moment we may be provided by the comparator network with assessment stating how close our situation resembles the previously seen situations. The key advantage of the comparator network is the fact that it is able to adapt to various levels and contexts that occur in data. In comparison with the other general, similarity-based approaches, such as Case Based Reasoning (CBR, see [1]), the one we propose is more flexible and provides more possibilities for fine-tuning the solution.

The approach based on comparators and comparator networks is a field-proven technology. Several uses in both Research and Development (R&D) and production software systems have been reported. The applications of comparator networks range from shape and character recognition in images to information retrieval in natural language text corpora (see [2]–[5]). A big incentive for using this technology is the existence of software library in Java that can be readily adapted for the needs of potential user. The library, as well as the accompanying services[1] are publicly available (see [6]).

II. Comparators

Compound object comparator is a construct dedicated to processing complex objects represented as data entities. We denote such a construct by com^{ref}. Such comparator can be identified with a function:

$$\mu_{com}^{ref} : X \times 2^{ref} \to [0,1]^{ref}, \tag{1}$$

where $X \subseteq U$ is a set of input objects to be compared and ref is a set of reference objects that we infer the similarity from. $[0,1]^{ref}$ denotes a space of vectors $\vec{v}$ of dimension $|ref|$, where each i-th coordinate in $v[i] \in [0,1]$ corresponds to an element $y_i \in ref$, $ref = \{y_1, ..., y_{|ref|}\}$. We will further call ref a *reference set*, while each $Y \subseteq ref$ will be referred to as *reference subset*. Additionally, $a(x)$ is a function that provides a representation of an object $x \in X$ w.r.t. a given attribute a. This representation is then used by the comparator while processing x. Similarly, each reference object $y \in Y$ is processed using its representation $a(y)$ for a given attribute a. If we are given an ordering on elements of reference set ref, i.e, $ref = \{y_1, \ldots, y_{|ref|}\}$ we can represent the function corresponding to the comparator as:

$$\mu_{com}^{ref}(x, Y) = Sh(F(\vec{v})), \tag{2}$$

where Sh is a (result) *sharpening function*, F is a function responsible for filtering the result before sharpening (defuzzyfication) and $\vec{v}$ is a vector of dimension equal to the cardinality

[1]Comparators at Dituel http://www.dituel.pl/Service,908/Comparators,1136/index.html

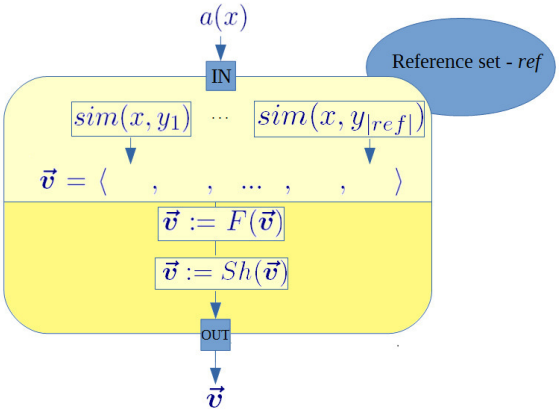


Fig. 1. Block diagram of the comparator's internal structure. The structure consists of two main layers: concurrent – responsible for the calculation of values for the initial proximity vector; sequential – responsible for processing the values from the previous stage. Blocks marked with *sim* are charged with calculation of unit similarity values. The internal structure of a such similarity block is described in Fig. 2 using UML.

of ref, composed of proximity (similarity, closeness) values between the object x and each of the reference objects in ref. Typically, F is based on combination of some standard, idempotent functions such as min, max, top, etc. [7] When Y is a proper subset of ref the positions in $\vec{v}$ corresponding to $y_i \notin Y$ are filled with zeros. Non-zero elements of $\vec{v}$ determine the degree of similarity (proximity) between the object x in question and each element of reference subset Y. Hence, the *proximity vector* can be defined as:

$$\vec{v}[i] = \begin{cases} 0 & y_i \notin Y \\ sim(x, y_i) & y_i \in Y \end{cases} \tag{3}$$

The value of similarity $sim(x, y)$ used above is calculated by means of a fuzzy relation [8] combined with additional mechanisms, as described in the next section.

III. Architecture of a Comparator

A comparator consists of two main layers (stages) – concurrent and sequential. The first of those layers is responsible for the calculation of values of proximity vector's coordinates for $x \in X$, $y_i \in ref$, with use of similarity function:

$$sim(x, y_i) = \begin{cases} 0: & Exc^{ref}_{Rules_i}(x) = 1 \vee y_i \notin Y \\ t_h(\mu(x, y_i)): & otherwise \end{cases}, \tag{4}$$

where $Y \subseteq ref$, t_h is a threshold function given as

$$t_h(z) = \begin{cases} 0 : & z<p \\ z : & z \geq p \end{cases}, p \in [0, 1]. \tag{5}$$

The value of p corresponds to the lowest acceptable similarity, μ is the basic similarity function, $Exc^{ref}_{Rules_i}$ is the function associated with prohibitive rules and i is an index of the coordinate of proximity vector for which the similarity is derived.

Similarities for different y_i can be calculated concurrently because they do not depend on each other. The internal, layered structure of a comparator is shown in Fig. 1. While each coordinate of the vector $\vec{v}$ is calculated independently, the final calculation of the value of the function (4) has to be performed in a sequence for a given pair of objects (x, y_i). This sequence of operations is illustrated in Fig. 2 in form of an UML activity diagram [9]. The processing in the first layer ceases when all coordinates of the proximity vector are derived. Only then can the comparator activate the next layer.

The processing in the comparator's second layer is performed sequentially. Hence, operations such as filtering by means of $F(\vec{v})$ and sharpening of proximity vector with $Sh(\vec{v})$ are performed one-by-one. This sequence yields the vector given by (3) as the final result.

If we take a wider look at the comparator for complex objects we may notice that if all the notions introduced above are composed, it can be expressed as:

$$\mu^{ref}_{com}(x, Y) = Sh(F(\langle sim(x, y_1), \ldots, sim(x, y_{|ref|})\rangle)) \tag{6}$$

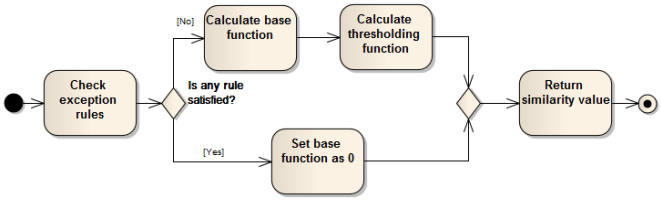


Fig. 2. An UML activity diagram of the *sim* block for a comparator calculating similarity between the input and reference object.

The notions introduced in sections II and III provide a systematic overview of the concept of comparator as a basic computing unit in systems aimed at supporting decisions, recognising patterns and at similar tasks. A single comparator (unit) is capable of measuring the level of similarity (proximity) of a given object to the reference set with regard to the attribute for which it was created. Comparator itself can be regarded as a decision support system, albeit a very rudimentary one. It can only handle very straightforward tasks associated with modelling a simple similarity. In order to use comparators in a more complicated situation one has to combine several of them into a network. Such a network is then an example of similarity-based inference system capable of modelling complex (similarity) relations occurring in data [2].

IV. Components of a Comparator Network

The operation of a comparator network can be interpreted as a calculation of a function:

$$\mu^{ref_{out}}_{net} : X \to [0, 1]^{|ref_{out}|}, \tag{7}$$

which takes the input object $x \in X$ as an argument and ref_{out} is a reference set for the network's output layer. The target set (codomain) of $\mu^{ref_{out}}_{net}$ is the space of proximity vectors. As in the previous situation, the proximity vector from the target space will be denoted by $\vec{v}$. Such a vector encapsulates information about similarities between a given input object x and objects from the reference set ref. Similarly to the case of

a single comparator, by ordering the reference set, i.e. taking $ref = \{y_1, ..., y_{|ref|}\}$, we get the value network's function of:

$$\mu_{net}^{ref}(x) = \langle SIM(x, y_1), ..., SIM(x, y_{|ref|})\rangle, \tag{8}$$

where $SIM(x, y_i)$ is the value of *global similarity* established by the network for an input object x and a reference object y_i. Global similarity depends on partial (local) similarities calculated by the elements of the network (unit comparators). Through application of aggregation and translation procedures at subsequent layers of the network these local similarities are ultimately leading to the global one.

A. Layers in Comparator Network

Each comparator network is composed of three types of layers: input, intermediate (hidden/internal) and output. A given network may have several internal layers [10]. Layer consists of comparators that are grouped together by the common purpose of processing a particular piece of information (attributes) about the object in question. Each layer contains a set of comparators working in parallel and a specific translating/aggregating mechanism. The translating and aggregating mechanisms are necessary to facilitate the flow of information (similarity vectors) between layers. As sets of comparators in a particular layer corresponds to a specific combination of attributes, the output of the previous layer has to be aggregated and translated to fit the requirements. This is done by elements called translators and aggregators, respectively. The translator converts comparator outputs to information about reference objects that would be useful for the next layer. The role of the aggregator is to choose the most likely outputs of the translator, in case there was any non-uniqueness in assigning information about input objects to comparators. The operation of a layer in the comparator network can be represented as a mapping:

$$\mu_{layer}^{ref} : X \to [0, 1]^{ref_l}, \tag{9}$$

where $x \in X$ is an input object and ref_l is the reference set for the layer.

Within a given layer only the local reference sets associated with comparators in that layer are used to establish (local) similarities. However, through aggregation and translation these local similarities become the material for synthesis of the output similarity and reference set for the layer. This synthesis is based on a translation matrix, as described in [11]. Function (9) is created as a superposition of: comparator's function (1), local (layer) aggregation function and translation. Local translation operation is responsible for filtering the locally aggregated results.

The input and internal (hidden) layers in the comparator network contain comparators with function (1) together with translators and local aggregators. The output layer contains the global aggregator responsible for returning the final result. The components of the comparator network are described briefly below. For details please refer to [11], [12].

B. Local Aggregator

Aggregators are a mandatory part of the network responsible for the synthesis of the results obtained by comparators. Aggregators are functions that operate on partial results of comparators. In the simplest case the network only needs a single *global aggregator* in the output layer. However, in the other network architectures it is included in other layers as well, in form of a local aggregator.

The local aggregator processes partial results of the network at the level of a given layer. The aggregator's operation depends on the type of reference objects and the output of comparators. It can be represented as:

$$f_{agg}^{ref_l} : [0, 1]^{ref_1} \times \ldots \times [0, 1]^{ref_k} \to [0, 1]^{ref_l}, \tag{10}$$

where k is the number of comparators in a given layer l, i.e. the number of inputs in the aggregating unit (local aggregator). ref_l is the output (resulting) reference set for layer l composed by means of the *composition rules* from the reference sets ref_i $(i = 1, \ldots, k)$ used by comparators in layer l.

C. Translator

The translator is a network component associated with the adaptation of results of one layer to the context of another layer (the one to be fed with). In other words, this element expresses the results of the previous layer (their reference objects) in reference objects of the current one. It uses reference objects of the next layer, taking into account the relationships between the objects of both layers [13]. The translator is defined by means of the translation matrix:

$$M_{ref_l}^{ref_k} = [m_{ij}], \tag{11}$$

where $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$ for m and n denoting cardinality of ref_k and ref_l, respectively. The matrix $M_{ref_l}^{ref_k}$ defines the mapping of objects in the set ref_k onto objects in the set ref_l. In practice ref_k is just a union of reference sets for all comparators in a given layer and ref_l is the target reference set. Values in the matrix are within $[0, 1]$.

D. Projection Module

This network unit appears in selected layers whenever there is a need for selecting a subset of coordinates (project the vector onto subspace) in proximity vector that will be further used in calculations. The selection of a particular coordinate may be based on its value (above/below threshold) and/or on the limitations regarding the number of coordinates that can be preserved. For the i-th coordinate in the proximity vector the projection can be the following:

$$\mu_{proj}(v[i]) = \begin{cases} \vec{\boldsymbol{v}}[i] & projection(\vec{\boldsymbol{v}}[i]) = 1 \\ 0 & projection(\vec{\boldsymbol{v}}[i]) = 0 \end{cases} \tag{12}$$

where $i \in \{1, \ldots, |ref|\}$ and $projection(a)$ for $a \in [0, 1]$ is a function of the form:

$$projection : [0, 1] \to \{0, 1\}, \tag{13}$$

The function $projection$ is the actual selecting mechanism. It decides whether a given coordinate is set at 0 or not. This function can be defined as a threshold, maximum, ranking function, etc.

E. Global Aggregator

The global aggregator is a compulsory element of the output layer. Unlike local aggregators, which process results within a single layer, the global one may process values resulting from all layers at the same time. In the simplified, homogeneous case, when all layers use exactly the same reference set, the global aggregator may be expressed by:

$$\mu_{agg}^{ref_{out}} : \left([0,1]^{ref}\right)^m \to [0,1]^{ref_{out}}, \tag{14}$$

where m is the number of **all** comparators in the networks, i.e. the number of inputs to the global aggregator.

In the more complicated, heterogeneous case, the sets in subsequent layers and comparators may differ. In this case the aggregator constructs the resulting (global) reference set ref_{out} in such a way that every element $y \in ref_{out}$ is decomposed into y_1 in reference set ref_1, y_2 in reference set ref_2 and so on, up to y_m in reference set ref_m. For a given input object $x \in X$ the value of similarity between x and each element of in ref_1, ...,ref_m is known, as this is the output of the corresponding comparator. To obtain the aggregated result we use:

$$\mu_{agg}^{ref_{out}} : [0,1]^{|ref_1|} \times \ldots \times [0,1]^{|ref_m|} \to [0,1]^{ref_{out}} \tag{15}$$

Note, that formula (15) is similar to the one for local aggregator (10). The essential difference is in the fact that the local aggregator is limited to a subset of comparators contained in a given layer, while the global one looks at all comparators in the network.

With all the definitions of units the comparator network can be expressed as a composition of mappings in subsequent layers:

$$\mu_{net}^{ref_{out}}(x) = \mu_{layer-out}^{ref_{out}}(\mu_{layer-int}^{ref_{k-1}} \ldots (\mu_{layer-in}^{ref_1}(x)) \ldots), \tag{16}$$

where ref_i stands for the reference set corresponding to layer i and ref_{out} is the reference set for the network as a whole. The general scheme of the comparator network is shown in figure 3.

V. Selected applications of comparator networks

Two examples of successful application of the comparator network methodology are presented. The two applications presented originate from rather unrelated areas. The purpose of showing them here is to highlight the versatility of comparator networks as a tool for dealing with real-life challenges. The first of the examples presented is associated with recognition and classification of texts in (quasi-)natural language. The second example relates to risk management in CBR systems [1]. The two examples presented are just a selection from a range of applications that were reported. For more examples refer to [2]–[5].

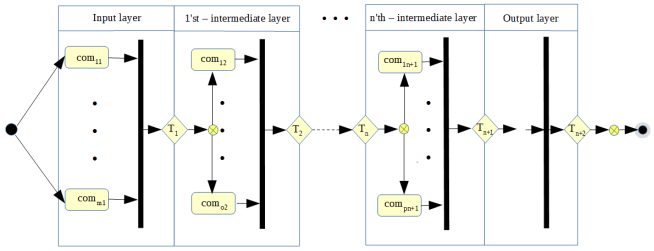


Fig. 3. General scheme of a comparator network in UML-like representation. Notation: com_{ji}- comparators, T_j- translators. Symbols: oval – comparator, thick vertical line – aggregator, rhombus – translator, encircled cross – projection module.

A. Classification of References in Scientific Articles

This application was conceived as an answer to a need defined in the course of the SYNAT project[2]. The overall goal of this project was to construct an integrated network platform for the storage, retrieval, management and delivery of digital information in areas of science and technology [14].

One of the tasks within the scope of SYNAT was to design and implement a sub-system for searching within repositories of scientific information (articles, biographical notes, etc.) using their semantic content (SONCA). The information contained in documents is by nature compound and relations discovered between their parts and other entities in data are crucial to preserve. One of the compound parts is the reference part in form of unstructured texts. The task is to identify the part of text that is a reference and recognise (name, classify) its parts, so that they can be given a semantical context which can greatly improve information retrieval and management. The result should be expressed by means of (sub-)objects classified to the already known classes such as authors, titles, journals, the publication date, and so on. For example, the following text found in the document: "*Sosnowski, L., Slezak, D.: Networks of Compound Object Comparators. In: Proc. of FUZZ-IEEE 2013 (2013)*" might be resolved by assigning the pattern ATRY to its parts, where: A stands for authors, T – title, R – reference volume (pRoceedings), and Y – the publication date (year).

TABLE I
Selected characteristics for similarities between the closest objects.

Measure	Value
Max	1.00
Min	0.50
Average	0.85
Median	0.87
Standard deviation	0.12
Variance	0.02

The whole process is divided into several stages: preprocessing, parsing and classification. The first stage is responsible for cleaning the data, sifting unnecessary data from the text

[2]SYNAT (www.synat.pl) was a major national R&D project financed by NCBiR as a part of the strategic research initiative.

TABLE II
TOP LEFT: BEST RESULTS FOR EXPERIMENT 1, TOP RIGHT: WORST RESULTS FOR EXPERIMENT 1, BOTTOM LEFT: BEST RESULTS FOR EXPERIMENT 2, BOTTOM RIGHT: WORST RESULTS FOR EXPERIMENT 2. NOTATION: P_1 - PRECISION FOR EXPERIMENT 1, R_1 - RECALL FOR EXPERIMENT 1, $F1_1$ - F1-SCORE FOR EXPERIMENT 1, P_2 - PRECISION FOR EXPERIMENT 2, R_2 - RECALL FOR EXPERIMENT 2, $F1_2$ - F1-SCORE FOR EXPERIMENT 2.

Pattern	P_1	R_1	$F1_1$	P_2	R_2	$F1_2$
ATR	1.00	1.00	1.00	0.75	1.00	0.86
RY	1.00	1.00	1.00	1.00	1.00	1.00
ATJVY	1.00	1.00	1.00	1.00	0.80	0.89
AT	1.00	0.98	0.99	0.61	0.92	0.73
ATRYP	1.00	0.93	0.96	1.00	0.73	0.83
ATVPYD	0.91	0.98	0.95	0.92	0.84	0.86
ATJVPYD	1.00	0.90	0.94	0.98	0.71	0.81
ATJVPY	1.00	0.86	0.92	1.00	0.58	0.72
AJVPYD	0.86	1.00	0.92	0.86	1.00	0.92
ATVPY	1.00	0.85	0.91	1.00	0.60	0.75

Pattern	P_1	R_1	$F1_1$	P_2	R_2	$F1_2$
ATC	0.33	0.67	0.44	0.00	0.00	0.00
ATYATYP	0.60	0.43	0.50	1.00	0.43	0.60
AYTRY	0.43	0.60	0.50	1.00	0.60	0.75
ATPYC	0.54	0.80	0.64	0.60	0.60	0.60
ATJVYP	1.00	0.50	0.65	0.50	0.25	0.33
ATVY	0.60	0.75	0.67	1.00	0.75	0.86
ATVYPR	0.56	0.83	0.67	1.00	0.50	0.67
ATJPY	0.94	0.58	0.70	0.89	0.60	0.71
ATRPYC	0.71	0.77	0.73	1.00	0.77	0.86
ATAT	0.57	1.00	0.73	0.00	0.00	0.00

Pattern	P_1	R_1	$F1_1$	P_2	R_2	$F1_2$
RY	1.00	1.00	1.00	1.00	1.00	1.00
ATRY	0.89	0.88	0.86	1.00	0.88	0.93
ATRPY	0.93	0.90	0.90	0.99	0.88	0.92
AJVPYD	0.86	1.00	0.92	0.86	1.00	0.92
ATRJPY	0.83	0.83	0.83	1.00	0.83	0.91
ATRPYCD	0.84	0.85	0.84	0.97	0.85	0.91
ATPYD	0.83	1.00	0.91	0.83	1.00	0.91
ATY	0.88	0.91	0.87	0.90	0.90	0.90
ATJVY	1.00	1.00	1.00	1.00	0.80	0.89
ATRP	1.00	0.75	0.86	0.88	0.88	0.88

Pattern	P_1	R_1	$F1_1$	P_2	R_2	$F1_2$
AYTJP	1.00	0.70	0.82	0.00	0.00	0.00
ATC	0.33	0.67	0.44	0.00	0.00	0.00
ATAT	0.57	1.00	0.73	0.00	0.00	0.00
ATJYR	1.00	0.60	0.75	0.00	0.00	0.00
AYT	0.94	0.87	0.87	0.13	0.13	0.13
ATJVYP	1.00	0.50	0.65	0.50	0.25	0.33
AYTP	0.73	0.85	0.76	0.37	0.30	0.33
ATYR	1.00	0.75	0.86	0.50	0.50	0.50
ATYP	0.83	0.69	0.75	0.61	0.47	0.53
ATPYC	0.54	0.80	0.64	0.60	0.60	0.60

structure and making it clear (e.g. lowercase, trimmed, etc.). It contains the prefix cleaning procedure, which eliminates unwanted prefixes e.g. "*[12] D., E., Willard, New trie data structures which support very fast search operations, Journal of Computer and System Sciences, v.28 n.3, p.379-394, June 1984 u00A0[doi>10.1016/0022-0000(84)90020-5]*" is an input text. The mentioned procedure cuts the part "*[12]*". The main idea of this stage is to dispose of all signs which can interrupt the further parsing or comparing process. The second operation is to replace all quotation marks which are not necessary but often interfere with the results.

The results reported below were obtained from two numerical experiments. For these experiments 400 examples of references (pieces of text) were drawn randomly from the corpus of scientific texts. The first experiment (Experiment 1) involved generating patterns for sub-objects (fragments of reference text) and composing (concatenating) them into a pattern for the whole reference. In the second experiment (Experiment 2) the pattern resulting from Experiment 1 was additionally compared with a reference set consisting of all possible patterns. The reference sets used were taken from the collection of already processed and classified articles in the SONCA subsystem [14]. Each of these reference objects had a clearly defined pattern it belonged to.

The experimental data were split into the training and testing parts containing 132 and 268 cases, respectively. The training sample was used to build the (local and global) reference set for the comparator network and for tuning of network's parameters. The test sample was used to assess the quality of classification (pattern identification). The experiment involved two main steps: (i) generation of reference patterns using similarities between sub-objects and (ii) comparison of generated patterns with the reference set of known (certain) patterns of bibliographic references.

Both experiments involved the use of a comparator network designed for this purpose. The network used local similarities to build (infer) the overall result. The particular network used had the most of the local similarity values at a relatively high level, within the $[0.8, 0.85]$ range. The other parameters of the network were set to default values. In particular, the threshold parameter p was set to middle (0.5) and the results were aggregated using simple averaging, i.e., assigning the same weights to all considered attributes. Table I presents certain statistics regarding the similarity calculated by the network.

The quality assessment was done with use of measures typical for classification applications such as *precision*, *recall* and *F1-score*. The solution (network) constructed achieved a global F1-score of 0.86 for Experiment 1 and 0.78 for Experiment 2. Both of these results can be considered sufficient for practical application. Table II contains more a detailed presentation of the best and worst results obtained during Experiments 1 and 2.

Patterns assigned to objects and patterns in the reference set were frequently quite complicated, which is a direct consequence of the almost non-existent format requirements and use of (quasi-)natural language. The table illustrates the kinds of patterns that appeared as a reference and were used to obtain results presented in table II. The types of elements of patterns corresponding to comparators in the network: Book (B), Country (C), DOI (D), Journal (J), Pages (P), Proceedings (R), Series (S), Title (T), Volume (V), Year (Y), Authors (A).

The experimental results presented show that the method used in Experiment 1 was more effective than the one used in Experiment 2. A closer look at the level of particular cases (objects) revealed that the reference set used in the second layer of the comparator network during Experiment 2 was

TABLE III
EXAMPLES OF OBJECTS IN NETWORK'S REFERENCE SETS.

Year
$START-CYYYY$
Jan $START-CYYYY$
Dec $START-CYYYY$

Pages
[p][0-9]{1,}-[0-9]{1,}
[p][\.][0-9]{1,}-[0-9]{1,}
[p][0-9]{1,}–[0-9]{1,}
[p][\.][0-9]{1,}–[0-9]{1,}

DOI
u00A0\[doi>.{0,}\]
u00A0\[doi>.{0,}

Authors
A. A. Abatan
A. A. Abonamah
A. A. Agrawal
C. W. Lin
F. Bonner

Volume
v\.\d+
v\d+\.
v\d+
v\.\d+ n\.\d+

Title
A comparison between conceptual clustering and conventional clustering
A Kalman-filter approach to equalization of CDMA downlink channels
A KDD System for the Discovery of Quantified Exception Rules

Journal
Fundamenta Informormaticae
IEEE transactions on computers
Journal of computer and system scien.

Proceedings
.{0,}\bproceedings\b \.{0,}
.{0,}\bproc\b\..{0,}
.{0,}\bproc\b\.\s.{0,}
.{0,}\bproc\b\s.0,

Country
POLAND
UNITED STATES
CHINA
INDIA

Structural patterns
ATJPY
ATRPY
ABTSVY
ATJY
ATJYP

insufficient. Some of the patterns that appeared in test data for Experiment 2 had no representative among the reference patterns, making comparison and identification invalid. There were also cases of errors in reference texts that were impossible to detect during preprocessing, but they were successfully filtered out by internal layers of the network, albeit at the cost of slightly reduced quality. Overall, the method based on the comparator network proved a useful addition to the system aimed at solving the task of processing texts in natural language.

B. Risk Assessment During Fire and Rescue Operations

1) Problem description: A Fire and Rescue (F&R) action is considered to be one of the most challenging environments for modeling and decision support. To date, there have been very few attempts to automate the decision making process in this area [15], [16], at least partially. One such attempt is the R&D project called ICRA (http://www.icra-project.org). The main goal of ICRA is to build a modern AI-based, risk-informed decision support system for the Incident Commander (IC), which improves situational awareness of the IC during F&R action, thus increasing the safety of firefighters. The basic ramifications and goals of the project can be found in [17].

One of the techniques introduced is the course of ICRA project is the *Threat Matrix*. The threat matrix groups the major types of threats with possible threat subjects. The threat matrix designed in the ICRA project is a significant extension of the model used currently by the German Fire Service. The new matrix is enriched with the possibility to define the degree (level) of the current threat which improves the representation of the current level of risks in the course of F&R operation. The extended version of the threat matrix is referred to as the *Risk Matrix*. The example of threats considered during the action using the ICRA version of the risk matrix is shown in table IV.

The experiments presented below were aimed at assessing the level of potential risks associated with the current operation through comparison (measuring similarity) with the previous, already fully described and classified cases retrieved from the repository. The underlying assumption of this approach is that the similarity between (circumstances of) two operations is an indicator of the similar types and levels of risks (threats) involved. In the operation scenario the IC would be informed

TABLE IV
RISK MATRIX REFLECTING OCCURRENCE OF THREATS WITH RESPECT TO THREAT TYPE AND POSSIBLE THREAT SUBJECTS. NOTATION: A1 - FEAR, A2 - TOXIC SMOKE, A3 - RADIATION, A4 - BURN-OUT, C - DANGEROUS CHEMICAL AGENT, E1 - COLLAPSE, E2 - ELECTROCUTION THREAT, E3 - DISEASE OR INJURY, E4 - EXPLOSION

Subject/Threat	A1	A2	A3	A4	C	E1	E2	E3	E4
People (ME)									
Animals (T)									
Environment (U)	-								
Property (S)	-	-				-	-	-	
Rescuers (MA)								-	
Equipment (G)	-	-						-	

that the current evolution of the situation shows similarities with a certain historical events and that some risks factors are more likely to emerge. As part of implementation of this experimental approach representation of an F&R operation as a vector of values have been devised.

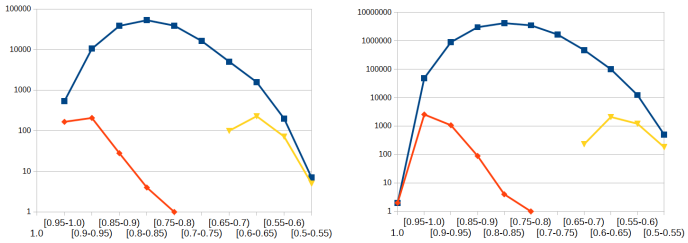


Fig. 4. Probability distributions for all, the best and the worst pairs of objects identified with the *leave-one-out* method. The graph on the left corresponds to results obtained from the examination of a set containing 406 operation reports. The graph on the right correspond to interventions. Both graphs use logarithmic scale. The Y axis shows the number of pairs while the X axis shows the clustered values of probability.

2) Experimental results: The proposed solution, based on comparator networks, was verified in the course of two experiments. The first of experiments was performed with use of a small data set of 406 F&R operation reports retrieved from the EWID[3] system. The experiment used the same data set, the same requirements and the same initial assumptions as the experiments reported in [18]. That allowed fully justified and

[3]Incident Data Reporting System used by the Polish State Fire Service

fair comparison of results obtained with methods used in [18] and those based on the comparator network.

The second experiment used data retrieved from the EWID as well. In this case the set contained 3736 reports. In both experiments the *leave-one-out* scheme was adopted resulting in, respectively, 164430 and 13953960 pairs of objects (reports) being considered in the calculation of similarities. It should be mentioned that the numbers of pairs that were actually used is lower than the cardinality of the respective cartesian products. This is due to filtering out those pairs that displayed very low similarity value, i.e. the pairs of very dissimilar reports.

The central tendency for results of both experiments is presented in table V. The distributions of similarity values obtained in experiments are shown in Fig. 4.

TABLE V
SELECTED MEASURES FOR CENTRAL TENDENCY FOR PAIRS OF OBJECTS: * - RESULTS FOR 406 CASES, ** - RESULTS FOR 3736 CASES. NOTATION: *all* - ALL PAIRS OF OBJECTS CONSIDERED, *bsp* - BEST SIMILARITY FOR INPUT PAIR, *wsp* - WORST SIMILARITY FOR INPUT PAIR.

Factor	*-all	*-bsp	*-wsp	**-all	**-bsp	**-wsp
Max	0.985	0.985	0.690	1.000	1.000	0.676
Min	0.530	0.791	0.530	0.477	0.797	0.477
Average	0.815	0.940	0.627	0.811	0.956	0.606
Median	0.820	0.946	0.630	0.824	0.961	0.608
Standard deviation	0.061	0.028	0.031	0.062	0.023	0.030
Variance	0.004	0.001	0.001	0.004	0.001	0.001

The results presented here for methods NoC $\frac{1}{n}$ and NoC WA were obtained with the default value of the threshold parameter $p = 0.5$. The weights for global aggregation used by NoC WA were calculated by means of the evolutionary algorithm. The best result was achieved for weights $\frac{2}{89}, \frac{60}{89}, \frac{1}{89}, \frac{26}{89}$ associated with, respectively, comparators for activities (sub-processes): notifications, disposals, recognitions, actions. These weights were obtained with the use of 33% of data (136 cases).

TABLE VI
COMPARISON OF VARIOUS METHODS FOR RECOGNITION OF RISKS. NOTATION: ESA - EXPLICIT SEMANTIC ANALYSIS, k-NN CANBERRA - k THE NEAREST NEIGHBORS WITH *Canberra* DISTANCE MEASURE [19], NOC $\frac{1}{n}$ - NETWORK OF COMPARATORS WITH AGGREGATION CALCULATED AS AVERAGE, NOC WA - NETWORK OF COMPARATORS WITH AGGREGATION CALCULATED AS WEIGHTED AVERAGE.

Method	Precision	Recall	F1-score
Naive Bayes	0.68	0.64	0.61
ESA	0.48	0.70	0.54
k-NN Canberra	0.74	0.74	0.69
NoC $\frac{1}{n}$	0.73	0.70	0.66
NoC WA	**0.79**	**0.75**	**0.71**

The final results for comparator networks were assessed in two steps. As part of the first step, the values of precision, recall and F1-score were calculated for every case of F&R operation in the data set and then averaged. These results are shown in table VI in comparison with the alternative classification methods applied to the same data. The second steps involved the analysis of the effectiveness of the method with respect to particular risks as identified in the risk matrix developed for the ICRA project. These results are presented in table VII.

TABLE VII
COMPARISON OF F1-SCORE VALUE FOR VARIOUS METHODS OF RECOGNITION WITH W.R.T. RISK TYPES. RISK TYPES ARE DEFINED TAKEN FROM RISK MATRIX [16]. NOTATION: NB- NAÏVE BAYES, ESA - EXPLICIT SEMANTIC ANALYSIS, k-NN C - k NEAREST NEIGHBORS WITH *Canberra* DISTANCE MEASURE [19], NOC $\frac{1}{n}$ - NETWORK OF COMPARATORS WITH AGGREGATION CALCULATED AS AVERAGE, NOC WA - NETWORK OF COMPARATORS WITH AGGREGATION CALCULATED AS WEIGHTED AVERAGE; * - RESULTS FOR 406 CASES, ** - RESULTS FOR 3736 CASES.

Risk	NB*	ESA*	k-NN C*	NoC $\frac{1}{n}$*	NoC WA*	NoC $\frac{1}{n}$ **
A1_MA	0.38	0.45	0.34	0.36	**0.39**	0.47
A1_ME	0.86	0.82	**0.91**	**0.91**	0.90	0.91
A1_T	–	0.07	0.09	0.12	**0.16**	0.14
A2_MA	0.81	0.84	**0.89**	0.85	0.88	0.70
A2_ME	0.83	0.84	**0.90**	0.89	0.89	0.84
A2_S	**0.29**	0.22	0.09	0.17	0.1	0.20
A2_T	0.05	0.14	0.09	0.13	**0.17**	0.17
A2_U	0.39	0.30	0.44	0.34	**0.45**	0.38
A4_G	–	0.08	0.21	0.11	**0.25**	0.14
A4_MA	0.30	0.22	0.35	0.24	**0.47**	0.34
A4_ME	0.27	0.17	**0.41**	0.16	0.36	0.33
A4_S	–	–	**0.40**	0.23	0.30	0.34
A4_T	–	0.13	–	–	**0.67**	0.06
E1_MA	–	0.11	**0.48**	0.22	0.42	0.40
E1_ME	–	–	**0.22**	–	–	0.21
E2_MA	0.11	**0.31**	0.17	0.07	0.24	0.30
E2_ME	–	**0.24**	0.20	0.09	0.14	0.24
E2_S	–	**0.15**	0.13	–	0.12	0.03
E3_G	–	0.12	**0.40**	–	–	0.08
E3_MA	–	**0.50**	0.16	0.13	0.17	0.28
E3_ME	–	–	0.12	**0.28**	0.13	0.42
E4_MA	–	–	–	–	**0.14**	0.13
E4_ME	–	–	–	–	–	0.14
E4_S	–	–	–	–	–	0.12

3) Interpretation of results: As shown in tables VI and VII, the best results were obtained for the *NoC WA* (Network of Comparators with Weighted Average) approach. In both experiments these assessments were reflected by values of standard quality measures [20], such as precision, recall and F1-score. The average value of precision for comparator-based approach was equal to 0.79 and by 0.05 higher than the average for the next best solution. A similar difference was observed for other measures (recall, F1-score) when comparing NoC WA with other classification methods. Detailed examination of the results w.r.t. particular types of risks, as shown in table VII, also favours the approach based on comparator networks. Overall, the NoC WA method proved to be a very promising approach to automation of decision support process for F&R operations [15], [16].

VI. SUMMARY OF THE POSITION

In the paper we put forward a novel methodology for reasoning as regards compound objects on the basis of their similarity to elements of knowledge base. This methodology offers a systematization of the process of constructing a system that identifies and classifies a compound object by comparing it to reference objects at various levels of abstraction. The proposed comparator networks complement and extend the existing gamut of similarity-based approaches to identification, management and classification of compound data entities.

An important advantage of the methodology presented in this paper is the availability of implementation. The implementation was done with domain experts in mind. For a user who understands the nature of the problem the materialization

of a solution in form of a comparator network comes quite naturally. The library `comparators-lib` that contains algorithms is implemented in JAVA [6]. It provides a collection of base classes and interfaces corresponding to constructs in the underlying methodology, such as a class for comparator network, class for layer, class for comparator, etc. The classes seamlessly facilitate data flow and parameter setting, so that the user can concentrate on designing the architecture of the solution and defining basic elements that lead to the final computational system. In this way, the whole process of building the comparator network for a given application becomes easier.

The current status of the technology based on the concept of comparators and comparator networks is presented in this paper. The functions and operation of underlying concepts and definitions were illustrated with examples of real-life applications. The next step for the development of this technology could be the creation of a networked platform providing access to methods and algorithms for comparator network architecture. Equipped with a well-designed graphical interface such a platform would greatly simplify the construction of network models for new applications. The platform would also provide a collection of templates that could be used for fast prototyping of the solution and finding optimal model parameters more efficiently. This can make it a useful tool for many different types of approaches based on similarity, i.e. semantic similarity [21], similarity-based reasoning, mereological similarity and approximate reasoning [22] as well as many others. All of them can be implemented by means of compound objects comparators.

REFERENCES

[1] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *Artificial Intelligence Communications*, vol. 7, no. 1, pp. 39–59, 1994. [Online]. Available: http://dl.acm.org/citation.cfm?id=196108.196115

[2] D. Ślęzak and Ł. Sosnowski, "SQL-based Compound Object Comparators: A Case Study of Images Stored in ICE," in *Proc. of FGIT-ASEA 2010*, ser. Communications in Computer and Information Science, vol. 117, 2010. doi: 10.1007/978-3-642-17578-7_30 pp. 303–316. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17578-7_30

[3] Ł. Sosnowski, "Characters recognition based on network of comparators," in *Techniki informacyjne teoria i zastosowania*, A. Myśliński, Ed. IBS PAN, 2012, vol. 4, pp. 123–134. ISBN 83-894-7555-3

[4] ——, "Inteligentne dopasowanie danych przy użyciu teorii zbiorów rozmytych w systemach przetwarzania danych," in *Analiza systemowa w finansach i zarządzaniu*, J. Hołubiec, Ed. IBS PAN, 2009, vol. 11, pp. 214–218. ISBN 9788389475220

[5] Ł. Sosnowski and D. Ślęzak, "How to design a network of comparators," in *Brain and Health Informatics*, ser. Lecture Notes in Computer Science, K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, and N. Zhong, Eds., vol. 8211. Springer, 2013. doi: 10.1007/978-3-319-02753-1_39. ISBN 978-3-319-02752-4 pp. 389–398. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-02753-1

[6] Ł. Sosnowski, "Framework of compound object comparators," *Intelligent Decision Technologies*, vol. 9, no. 4, pp. 343–363, 2015. doi: 10.3233/IDT-140229. [Online]. Available: http://dx.doi.org/10.3233/IDT-140229

[7] R. W. Quackenbush, "On the composition of idempotent functions," *algebra universalis*, vol. 1, no. 1, pp. 7–12, 1971. doi: 10.1007/BF02944949. [Online]. Available: http://dx.doi.org/10.1007/BF02944949

[8] J. Kacprzyk, *Multistage Fuzzy Control: A Model-based Approach to Fuzzy Control and Decision Making*. John Wiley & Sons, 2012. ISBN 9780470744161

[9] J. Rumbaugh, I. Jacobson, and G. Booch, *The Unified Modeling Language Reference Manual, 2nd Edition*. Pearson Higher Education, 2004. ISBN 0321245628

[10] Ł. Sosnowski and D. Ślęzak, "Networks of compound object comparators," in *FUZZ-IEEE*. IEEE, 2013. doi: 10.1109/FUZZ-IEEE.2013.6622547. ISBN 978-1-4799-0020-6 pp. 1–8. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/FUZZ-IEEE.2013.6622547

[11] ——, "Fuzzy set interpretation of comparator networks," in *Pattern Recognition and Machine Intelligence - 6th International Conference, PReMI 2015, Warsaw, Poland, June 30 - July 3, 2015, Proceedings*, 2015. doi: 10.1007/978-3-319-19941-2_33 pp. 345–353. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19941-2_33

[12] Ł. Sosnowski, A. Pietruszka, and S. Łazowy, "Election algorithms applied to the global aggregation in networks of comparators," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: 10.15439/2014F494 pp. pages 135–144. [Online]. Available: http://dx.doi.org/10.15439/2014F494

[13] Ł. Sosnowski, "Applications of comparators in data processing systems," *Technical Transactions, Automatic Control*, pp. 81–98, 2013.

[14] R. Bembenik, Ł. Skonieczny, H. Rybiński, and M. Niezgódka, Eds., *Intelligent Tools for Building a Scientific Information Platform*, ser. Studies in Computational Intelligence. Springer, 2012, vol. 390. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-24809-2

[15] L. Han *et al.*, "Firegrid: An e-infrastructure for next-generation emergency response support," *Journal of Parallel and Distributed Computing*, vol. 70, no. 11, pp. 1128 – 1141, 2010. doi: http://dx.doi.org/10.1016/j.jpdc.2010.06.005. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2010.06.005

[16] A. Krasuski, A. Jankowski, A. Skowron, and D. Ślęzak, "From sensory data to decision making: A perspective on supporting a fire commander," in *Web Intelligence/IAT Workshops*. IEEE Computer Society, 2013, pp. 229–236. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/WI-IAT.2013.188

[17] Ł. Sosnowski, A. Pietruszka, A. Krasuski, and A. Janusz, "A resemblance based approach for recognition of risks at a fire ground," in *Active Media Technology - 10th International Conference, AMT 2014, Warsaw, Poland, August 11-14, 2014. Proceedings*, 2014. doi: 10.1007/978-3-319-09912-5_47 pp. 559–570. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09912-5_47

[18] A. Krasuski and A. Janusz, "Semantic tagging of heterogeneous data: Labeling fire & rescue incidents with threats," in *FedCSIS*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2013, pp. 77–82. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6643979

[19] F. Malik and B. Baharudin, "Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the {DCT} domain," *Journal of King Saud University - Computer and Information Sciences*, vol. 25, no. 2, pp. 207 – 218, 2013. doi: http://dx.doi.org/10.1016/j.jksuci.2012.11.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1319157812000444

[20] A. M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the web," *ACM Transactions on Internet Technology*, vol. 9, pp. 10:1–10:24, July 2009. doi: http://doi.acm.org/10.1145/1552291.1552293. [Online]. Available: http://doi.acm.org/10.1145/1552291.1552293

[21] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *CoRR*, vol. abs/1105.5444, 2011. doi: http://dx.doi.org/10.1613/jair.514. [Online]. Available: http://arxiv.org/abs/1105.5444

[22] L. Polkowski, *Approximate Reasoning by Parts: An Introduction to Rough Mereology*, ser. Intelligent Systems Reference Library. Springer, 2011. ISBN 9783642222795

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 41–48
DOI: 10.15439/2016F594
ACSIS, Vol. 9. ISSN 2300-5963

DAILY TOURISTIC PLAN RECOMMENDATION USING TEXT MINING

Kerem Turgutlu
Istanbul Technical University
Email: keremturgutlu@gmail.com

Erkan Isikli, PhD
Istanbul Technical University
Email: isiklie@itu.edu.tr

"This study focuses on the proposal of a recommender system for daily touristic plans. In order to construct such a system it is further examined that there is a need of text mining applications. Moreover, Sentiment Analysis and Keyword Extraction techniques are evaluated by developing and testing different approaches. Sentiment Analysis approaches are examined step-by-step in order to pick the best among them to score restaurant data. Similarly, Keyword Extraction is evaluated from various perspectives of statistics, visualization and machine learning. By the end of the paper the structure and the flow of the proposed system is illustrated upon the chosen approaches which were tested throughout this paper."

I. Introduction

THE main aim of this study is to create a dynamic environment that enables users to find the best activities and dining options on real time around a desired travel destination. The project tackles certain problems that many travelers face in a day-to-day basis such as wasting unnecessary amount of time planning what to do in a touristic area and struggling with the difficulty of finding activities or dining places which suits them best. Inadequacy of conventional travel websites, blogs, reviews and their lack of simplicity due to large amount of data makes such decisions time consuming.

In order to create such a system which allows users to pick their interests from a text cloud and later allows them to choose from the recommendations of auto-generated day plans considering their needs, desires and budgets, one can make use of two major text mining techniques: *Sentiment Analysis* and *Keyword Extraction*.

The scope of this idea of creating a smart online travel day recommendation system is considerably wide as there are thousands of travel destinations worldwide. For simplicity and testing matters, only the restaurants in Amsterdam area are considered at the modeling stage, but it should be noted that each and every step of this study can be applied to any touristic attraction in every travel destination as long as there is available data online. All datasets are generated by collecting user reviews from TripAdvisor due its simplicity of API integration and cost efficiency. Additionally, later during the product development stage of the proposed recommender system, many other review sources will be combined to elaborate the findings [1].

II. Text Mining

The "Text Mining" is often generalized as processing structured or unstructured but information holding data for the sake of generating patterns of information [12]. Natural Language Text is the major study and analysis source of Text Mining, thus exponentially growing web-based textual information makes it more attractive every passing year. There are various ways of analyzing textual documents via text mining, as well as many types of information gathered by using its techniques.

For both data and text mining it is primarily important to analyze or process a data which potentially holds information. In other words the actions taken by data and text mining tools should be in a way that it illustrates or generates an information from a given data. Aside from this mutual need of having and analyzing potentially useful data; text and data mining differ substantially when it comes to type of data that is used. Data mining deals with incomprehensible data with binary, nominal, ordinal and interval features which only deploy a meaning when certain algorithms are used or statistical interpretations are made. On the other hand, text mining data is already comprehensible and gives a textual information even without processing it. This uniqueness of explicit information bearing puts text mining one step ahead [12]. Nonetheless, for both cases detecting an informative pattern is equally tricky since both data mining features and text mining data are incomprehensible to computers or machines which tries to interpret them.

This uniqueness of text mining is the reasons that it is adopted at the core of this project. Many traveling sites and other sources offer direct information about restaurants or touristic attractions but this information generally does not go beyond cost, address, opening-closing hours and other type of strict data. Most of the time what real travelers seek during their explorations are fast recommendations which suits them best usually from their close network of friends and families. Our aim goes beyond the limited circle of friends and families, considers millions of available reviews.

III. Sentiment Analysis

Sentiment analysis or opinion mining is the task of identifying the subjectivity of a document and later determining its class as being; neutral, positive or negative.

Sentiment analysis is widely used in business related domains such as; marketing, customer satisfaction and benchmarking, as well as in political science, law, sociology and psychology [18].

Statistics or machine learning algorithms are used to classify the documents' sentiment. Even though state-of-art methods and algorithms tend to give satisfying results, sentiment analysis is a difficult task when the complexity of people expression, unrelated lexical content, negations and rhetorical devices as irony and sarcasm is considered [18]. Deciding positivity and negativity of a text may come out differently with errors even when it is done manually by 2 different human candidates. This shows the complexity of human expressions and perceptions. As the complexity of the domain and opinion increases the more difficult the task becomes. Relatively sentiment analysis on a product is easier to a political opinion [18]. Moreover in this paper sentiment analysis on restaurants is studied, using machine learning methods and probabilistic approaches.

Particularly in this paper, the main objective had been to help tourists to be able to have the optimal day plan with significant amount of time saving. In order to guarantee them to have the best possible experience: each restaurant and touristic attraction should be scored. Different modeling approaches are studied and evaluated step-by-step. Moreover, tourist reviews are collected from *TripAdvisor* and processed by applying sentiment analysis using different methods and later picking the dominating method to score each restaurant and activity.

IV. SENTIMENT ANALYSIS APPLICATION

A. TF-IDF Approach I

First part of the sentiment analysis is data collection. As mentioned earlier, the whole data (training and test together) is collected from TripAdvisor website under Amsterdam search. In the first model, only the 50 top-ranked restaurant reviews are used. The whole data consists of 11688 reviews in total, which are scraped by using Kimono API creator [11]. The features that are collected to be used in modeling are: *title of the review*, *review text*, *number of stars out of 5*, *total number of previous reviews made by the same user* and *the total number of helpful reviews that are made by the same user so far*. The last two features are not directly related with the sentiment analysis; in fact, they are not used in application, but they might be helpful to assign weights on each review once they are tagged as a positive or a negative review and may have an effect on the overall score of a restaurant. In total, there are 11027 positive and 661 negative reviews combined.

In this part, supervised learning techniques are employed and in order to apply these techniques effectively, labels are assigned as 1 for a positive review and 0 for a negative review. Many studies in the related literature suggest the use of an additional and a priori labeling which conditions on subjectivity or objectivity of the text. However, since we are dealing with customer reviews, we are almost certain that all the reviews are subjective at some point and does not state facts as in news articles or product descriptions. Hence, it would be convenient to bypass subjectivity analysis part assuming all the reviews are subjective and only use a binary labeling for the supervised learning. To do so, a simple code was written to automatically separate the reviews with less than 4 stars as negative reviews, from the positive reviews (those with 4 or higher stars). The title and review text were then concatenated as a single string. Finalizing the data set into two features as *full text* which has the text and as *sentiment* which contains the class labels 1 for a positive review and 0 for a negative review. Tail of the structured data is shown in Table 1.

TABLE I.
TAIL OF THE STRUCTURED DATA WITH FULL TEXT AND SENTIMENT FEATURES

	Full Text	**Sentiment**
11683	Worth every cent!. Went a little out of the ce...	1
11684	Very good experience.. We had a amazing night …	1
11685	EXCELLENT MEAL. Have been here before two year…	1
11686	Amazing. Went to this restaurant for boyfriend…	1
11687	Absolutely stunning throughout. Restaurant was..	1

In the next step before training the data set the text is preprocessed in order to get rid of the punctuation marks, to get rid of html markups, to deal with emoticons and with lowercase letters [7]. Even though punctuation marks address significance about the sentiment class identification it may lead the classifier into an unwanted direction, in which case "!" may be a negative or positive claimer [7]. In almost every NLP (Natural Language Processing) tasks tokenization is necessary and for this study each individual sentence is broken into words for further processing [6]. In the next step the reviews are converted into a feature matrix consisting rows for reviews and columns for each tokenized words. In this study features are single words - unigrams, but different n-grams could be chosen for different processing purposes [6].

After tokenization step, term frequencies of each single unigram-feature under each document is denoted by tf (t, d). An illustrative example 3 different document sentences is given in Figure 1.

{This: 1, restaurant: 2, is: 3, good: 4, bad: 5, ok: 6}

D1: [1, 1, 1, 1, 0, 0]
D2: [1, 1, 1, 0, 1, 0]
D3: [1, 1, 1, 0, 0, 1]

Fig 1. Each document is given with its feature matrix *tf (t, d)*.

The sentences in Figure 1 are read in this order: “This restaurant is good”, “This restaurant is bad” and “This restaurant is ok”.

Unimportant English words are overly weighted when dealing with term frequencies. In order the tackle this problem and give each feature a weight corresponding to its importance , a special form of feature vectorizer called *term frequency – inverse document frequency tf-idf (t, d)* is used [7, 8].

$$tf - idf(t,d) = tf(t,d) \times idf(t,d)$$

The inverse document frequency is calculated as follows:

$$idf(t,d) = \log \frac{n_d}{1 + df(d,t)}$$

In the last equation, n_d is the total number of documents and $df(d,t)$ is the number of documents that contain term *t*. Addition of term 1 in the denominator allows smoothing and deals with log expression [9].

Before further processing, the data is split into training and test sets with 0.67 to 0.33 ratio. Later, in Python a pipeline is constructed including a tf-idf vectorizer and a Logistic Regression classifier with L2 regularization and parameter C = 10 [7]. It should be noted that once the documents are converted into a feature matrix, any classifier might have been used. SVM, MaxEnt, Random Forests are commonly used classifiers for this purpose [2].

There are various machine learning metrics used to analyze how well a model performs: accuracy, precision, recall, and F measures [10]. Since the test set is pre-labeled as in every supervised machine learning model, a comparison between true classification labels and predictions could be made. To provide a better understanding, a confusion matrix is given in table II and related metric functions are provided.

TABLE II.
CONFUSION MATRIX

	Predicted NO	**Predicted YES**
Actual NO	True Negative	False Positive
Actual YES	False Negative	True Positive

Accuracy = TN + TP/ (TN + TP + FP + FN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1 = 2 * (Precision * Recall) / (Precision + Recall)

The predictive value negative, which can be given by:

$$\frac{\textbf{True Negative}}{\textbf{True Negative} + \textbf{False Negative}}$$

This metric addresses the power of predicting negative reviews. Due to lack of negative reviews, under fitting occurred for this specific case. In the proceeding, more negative reviews are collected to improve predicting the negative reviews as well.

B. TF-IDF Approach II

After additional data collection, the number of negative sentiment reviews increased by 2100 and the model trained with the same parameters but with a larger data set.

Predictive value negative metric increased by more than 0.40 points as we anticipated and obtained as 0.7452. Nonetheless, this increment is not a pure improvement since the newly generated test data is different from the previous dataset both in scale and observations. Thus, in order to evaluate the performance in an unbiased way, the same splits from the old data were created and tested with the new model.

The performance of tf-idf model is increased by collecting more negative reviews, especially significantly in the power of predicting negative reviews with a point difference more than 0.30.

C. Boolean Multinomial Naive Bayes Approach I

According to many studies in the related literature, when it comes to sentiment analysis, Naive Bayes tends to give promising results. This approach has Bayes Theorem in its core and naive term comes from its simplicity due to the avoidance of the dependency of occurrence of each word. It is assumed that each word is independent [3 4 5]. The theorem basically argues that the class or the sentiment of a document is the maximum probability it gets, given such a document: sequence of words. The documents we are referring throughout this part are individual tourist reviews. To briefly illustrate steps of Naïve Bayes mathematically [3 4 6]:

C: Class, D: Document

1) Objective Function: $argmax[P(C|D)] \; \forall \; C$
2) Expand $P(C|D)$:
$$\frac{[P(D|C) * P(C)]}{P(D)}$$
$P(D)$ is a global constant. So, the numerator part is enough.
3)
$$[P(D|C) * P(C)]$$
$P(C)$: Proportion of class C upon all documents.
4) Expand $P(D|C)$:

D: documents composed of unigram tokens.
Represented as:

$$P(w_1, w_2, .., w_n|C)$$

$$w_n: nth\ word\ in\ document$$

Word occurrences are considered independent:

$$P(w_1|C) * P(w_2|C) *..* P(w_n|C)$$

5) Final estimator:

$$\mathrm{P(C)} * \prod_{\mathrm{i}} \mathrm{P}(w_i|\mathrm{C})$$

Log scaling is used in order to prevent floating points and to prevent excessive weights on frequently used words.

$$\mathrm{argmax}[\log(\mathrm{P(C)}) + \sum_{\mathrm{i}} \log(\mathrm{P}(w_i|\mathrm{C}))]$$

Boolean Multinomial Naive Bayes is a special case of Naive Bayes with steps:

1) Preprocess text.
2) Remove all duplicate words in each document.
3) Do Naive Bayes.

$$P(w_i, C) = (count(w_i, C) + 1)/(count(C) + |V|)$$

|V|: vocabulary size

+1 and +|V| is for Laplace smoothing in order to avoid none observations.

Another problem that this method faces is the biased weighting due to count differences of each word in each class. To get over this mighty problem, likelihood approach is defined [6]:

$$\mathrm{P(w|C)} = \mathrm{f(w, C)}/\sum_{\mathrm{C}} \mathrm{f(w, C)}$$

f: frequency of word w in class C.

Again the same dataset from TFIDF Part 2 is used but with a different sample size. In order to balance weights of each class and their effects on prediction, negative and positive review sample sizes are selected equally.
As a new data set; 2100 negative and 2100 positive reviews were merged. First the "stop words" in English language are removed. Then it is split into train and test data by 80% to 20% [7].

In the preprocessing step, regular expressions and stop words are removed and all the words are re-written in lower case. Next, sentences are broken into unigram word tokens. Finally in order to begin Boolean Multinomial Naive Bayes each duplicate of words in all sentences are removed.

An Out-of-Vocabulary (OOV) word is considered to appear equally likely on both class such as words like "restaurant", "food", etc. Thus, assigning a 50% likelihood to those unseen words when dealing with smoothing is important in order to avoid undefined operations like 0/0, log (0) or to avoid smashing all the chain probabilities to 0. This modification allows the algorithm to deal with unfamiliar and unrelated words equally [6].

Smoothing may be changed and tested accordingly to evaluations of the study. It is decided to assign 0.5 to likelihood for OOV words and to assign a very small number close to 0 such as 1e-15 for log (0) cases. Basically; given a word if that word is not in the given class but in the other class: assign 1e-15. If the given word is not appearing on both classes then assign 0.5.

Collective frequencies are simply the sum of the frequencies in both classes. Taking a word token into account. If it is not found in the negative class then it is set and labeled as "no found", in smoothing step that expression is replaced with the minimum value 1e-15.

Naive Bayes has improved the previous model's power of predicting sentiment for class 0 (negative) reviews. It used be around 0.6460 and after Naive Bayes Classifier it increased to 0.7845 boosting it up almost 0.15 points.

In general, Naive Bayes gives high hopes on both predicting positive and negative sentiment equally. Performance evaluations can be improved by collecting more data, improving preprocessing steps and applying fine tuning by using stratified k-fold [7].

D. Boolean Multinomial Naive Bayes Approach II

In this part of the Boolean Multinomial Naive Bayes, rather than using equal sized samples for both classes all of the data is used for better fitting. Bad/Good Ratio for all data is around 0.1788. Train and Test data is split by 80% to 20%. Further for better validation of the model, stratified k-fold cross validation is used where k = 5. By using cross validation training data is split into 5 different train and validation parts for best selection, while the bad/good ratio is preserved [7].

Using a data set composed of 1341 negative reviews and 7517 positive reviews was not enough explanatory when it comes to predict negative reviews. Previous part of this model outperformed this case. Equal sized samples with good/bad = 1 ratios tend to fit negative reviews better where this case with large positive dataset and low negative dataset performed only as good as fair coin toss. Here, the major problem is the convergence of the predicted sentiment of the reviews to a single class due to its higher weight and feature scale compared to the other class. As positive training samples dominate in size and negative samples lack to describe an unseen test sample; the minority class tend to converge into the dominating class. Since it does not represents all the features or word tokens.

In literature and from the previous parts of study more data collection can be considered as a better approach. But, at the same time enough descriptive information for each class should be collected in order to perform equally good in each case.

One might argue that Naive Bayes Algorithm can deal with unequal sized class samples by having P(C) class probabilities inside the formula [3, 4, 6]. Even though this is the case sometimes P(C) ratios can weight more than word frequency ratios.

For example, a sample review is examined:

Word “great” is a feature which should be considered as a positive identifier according to human logic and English language. But the following frequency values pulled from the model indicates the effect of “great” on prediction is not great as the class frequency itself, hence not making the impact it should have made.

Word “great”:

Negative frequency: 0.004
Positive frequency: 0.014

“great positive frequency”/ “great negative frequency” = 3.5

“positive class frequency”/ “negative class frequency” = 5.6

Having 5.6 > 3.5 and also seen from this example, in some cases class ratios have more weights on the predictive algorithm than the actual descriptive words. So to finalize, for cases like restaurant review sentiment analysis with 2 classes it can be considered as a better approach to have balanced class samples. Additionally even if equal size samples are not favorable in some cases, there should be a threshold value to prevent overweight of a class frequency rather than the actual words.

Table below represents the model evaluations for each approach used in sentiment analysis case.

TABLE III
PERFORMANCE METRICS

	Accuracy	Precision	Recall	F1 Score	Predictive Value Negative
TF-IDF I	0.9567	0.9799	0.9953	0.9774	0.3363
TF-IDF II (More Data)	0.945	0.9766	0.9802	0.9681	0.7452
TF-IDF II (Same Old Test Data)	**0.9653**	**0.9886**	**0.9851**	**0.9816**	0.646
Bool. Multi. Naïve Bayes I	0.8155	0.8308	0.7845	0.807	**0.7845**
Bool. Multi. Naïve Bayes II	0.8985	0.9188	0.9655	0.9465	0.5283

V. RESTAURANT SENTIMENT SCORING

In overall, Multinomial Naïve Bayes Part 1 results outperformed other methods and in this section each individual restaurant review is predicted by that same classifier. Later positive and negative reviews of each restaurant are used in order to score them to create a ranked list of restaurants. Percentage of positive reviews are given to assign a score to restaurants. A short sample list of scored restaurants are given in Table IV as an illustration. The scores will allow the proposed recommender system to recommend top restaurants which are related with the user’s interests.

TABLE IIV
RESTAURANT SCORES

Restaurant	Score
Arendsnest Dutch Beer Bar	0.8419
Bakers & Roasters	0.7663
Biercafe Gollem	0.8424
Bird Thai Snackbar	0.8242
Bord'Eau	0.9427
Brasserie Ambassad	0.8533
Brasserie SenT	0.7642
Brasserie Vlaming - Amsterdam	0.84
Braziliaans Grill Restaurant Rodizio.nl	0.2241
Broodje Bert	0.7895

VI. KEYWORD EXTRACTION

Information Extraction (IE) is widely used in order to get a smaller structured information from a document by using statistical analysis, machine learning and NLP (Natural Language Processing) techniques. IE also contains sub-tasks as NER (Named Entity Recognition), Semi-Structured IE, Terminology Extraction, Keyword Extraction and Audio Extraction. Interestingly, IE is not only used to extract from textual data but also involves studies in multimedia extraction. There are mainly three widely accepted methods to extract information from a document, which are Hand Written Regular Expressions, Classifiers as Naive Bayes and MaxEnt and finally Sequence Models as Markov Models or Conditional Random Fields. [23].

In this paper, keyword extraction plays a vital role on determining the characteristics of a particular restaurant and in the future implementations to determine the characteristics of a touristic activity. The main objective of extracting keywords or information from a particular domain is to find out which attributes describes that entity best in an optimal and efficient manner. Later these extracted keywords will be clustered into groups according their similarity in order to generate a word cloud. This word cloud or network of descriptive words will be represented to the users allowing them to pick words according to their interests. Word selection phase will later lead to generate an

optimal day plan by co-working with the scores we obtained in the sentiment analysis phase.

Three different approaches are tested during keyword extraction task. These approaches are Rake Algorithm, KeyGraph Algorithm and machine learning approach by Random Forests respectively. Again, for this part the same restaurant data set from *TripAdvisor* is used for test and evaluation purposes. Additionaly, following applications are modeled and tested by datasets generated from *Arendsnest Dutch Beer Bar* only.

VII. KEYWORD EXTRACTION APPLICATION

A. RAKE Algorithm

Rapid Automatic Keyword Extraction (RAKE) Algorithm is the first approach adopted. Its good statistical interpretation and computational effectiveness due to its fast nature makes it a desirable candidate.

RAKE involves the following steps:

1. Data Preparation and Processing
2. Candidate Keywords Generation
3. Keywords Selection

In data preparation step, tourist reviews of Arendsnest Dutch Beer Bar are concatenated in order to form a single document. This document holds the latest 215 reviews that are made recently. Later, in the preprocessing step regular expressions, stop words, html markups and emoticons are removed from the document, also all words are lowered.

Candidate keywords are the tokens we would like algorithm to statistically evaluate and later assign a desirable amount of them as true keywords. So, a smart interpretation should be made while generating candidate keywords from a document. Also, one must have a good prior knowledge about the domain which of the document which will be processed. Here, we are dealing with restaurants and in contrast to a news article or a scientific paper keywords would not be longer than three words. Most of the time a restaurant would be described by its cuisine or atmosphere. Hence, candidate keywords are appended to a list by generating 1, 2 and 3 – gram tokens. Each n-gram group later evaluated separately in order to avoid cross dominations across groups.

In the final step, two different methods are employed in order to select keywords from n-grams. Each method use a statistical interpretation in order to score candidates and later outputs the desired amount of keywords or the ones that are above a given threshold.

First method is Best Match (BM):

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i)\left[\frac{f(q_i,D)*(k_1+1)}{f(q_i,D)+ k_1*(1-b+b*\frac{|D|}{avgdl})}\right]$$

|D| = Document length in tokens
avgdl = Average document length in tokens
f (qi, D) = Frequency of candidate q in document D
k1, b = Free default parameters

Due to lack of an error function and thus lack of an optimization, default values are given as k1 = [1.2, 2] and b = 0.75. IDF part is ignored in calculation since the stop-list the common English words in the preprocess step. Also having a single document gave identical scores for every k value in the range. Top 5 keywords obtained by BM for each n-gram group is given in table V.

TABLE V
TOP 5 KEYWORDS OBTAINED BY BEST MATCH

Top 5		
1-gram	2-gram	3-gram
"beers"	"dutch beers"	"great beer selection"
"selection"	"dutch beer"	"dutch craft beer"
"place"	"beer selection"	"dutch beer bar"
"staff"	"great beer"	"great beer bar"
"friendly"	"beer bar"	"best beer bar"

Second method is TF-IDF:

$$score = \left(0.5 + 0.5 * \frac{f(t,d)}{max_t f(t,d)}\right) * \log(\frac{N}{n_t})$$

f (t, d) = frequency of term t in document d
N = number of documents
n_t = number of documents containing term t

Logarithmic part of the score equation represents the IDF and it is neglected since we are processing a single document. As a result after applying TF-IDF method, the same top 5 keywords obtained for every n-gram group as in the BM method. Both TF-IDF and BM gave identical results with one documents case. Our conclusion here is that term frequencies of keyword tokens are highly correlated with the fact of being a keyword. Term frequency is later adopted as an important feature in machine learning model approach. RAKE is an incomplex and fast algorithm which yields satisfying results.

B. Key-Graph Algorithm

Key-Graph Algorithm is a visual indexing tool that is used to represent the characteristics of a single document. The algorithm creates a visual map containing clusters of words according their frequencies and co-occurrences [24].

Key-Graph involves the steps below:

1. Data Preparation and Preprocessing.
2. Extracting Foundations.
3. Extracting Columns.
4. Extracting Roofs.

In the first step, same preparation and preprocessing is applied and additionally each word is stemmed by Porter Stemmer algorithm. Different from grouping keyword candidates as n-grams in Key-Graph the aim is to find long keywords. So a candidate keyword list is created by deriving 2 and 1- gram tokens from 3-grams. Later, candidate phrase list is sorted by their frequencies in decreasing order. As it is mentioned above a relatively longer keyword is more

favorable to others in this algorithm. So, if 1 or 2-gram phrases have the same frequency as their parent 3-gram phrase, the algorithm automatically eliminates lower gram phrases from the candidate list. After this elimination step the sorted candidate phrase list is finalized. Top 50 candidates is chosen empirically for further steps.

Next, the association or co-occurrence scores of word pairs from the top 50 list is computed in order to cluster them by their scores.

$$assoc(w_i, w_j) = \sum_{s \in D} \min(|w_i|, |w_j|)$$

|w| = word count.
s = Sentence, a single tourist review in our case.
D = Document, collection of reviews of a restaurant.

Below is the pair association scores of pairs from our top 50 list. The pairs which are not included here have 0 association scores.

Words are clustered according to their scores as shown in figure 1 below. This is a multiply connected graph with two clusters, top group has association score of 2 and the other has score of 1. Dashed line connects two groups. Since each group has equal level of connection within themselves, 2!*3! = 12 different combinations can be maintained.

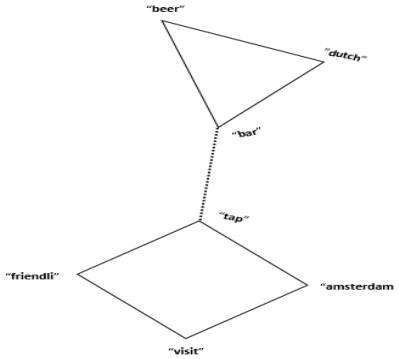


Fig 1. Extracted Foundations

Next, in order to extract columns to be added to foundation two functions are defined:

$$based(w, g) = \sum_{s \in D} |w|_s * |g - w|_s$$

$$neighbors(w, g) = \sum_{s \in D} \sum_{w \in s} |w|_s * |g - w|_s$$

w = words in top 50 list excluding the words in clusters.
g_i = graph including words in ith cluster

By based and neighbor scores key values of each word w is to be calculated as follows:

$$key(w) = 1 - \prod_{g \in G} \left(1 - \frac{based(w, g)}{neighbors(w, g)}\right)$$

Later top 5 ranking words w selected empirically according to their key scores. This key score represents the closeness to a cluster and to a specific word in that cluster. Words selected to be added to clusters as columns: [u'great', u'select', u'staff', u'place', u'tri']

These words are paired with words in graph clusters and scored according to column scoring function:

$$column(w_i, w_j) = \sum_{s \in D} \min(|w_i|_s, |w_j|_s)$$

Each column pair score resulted in 0 indicating no new term to be added next words in any of the clusters. Our final graph is the one we obtained in previous steps shown in fig 1.

C. Machine Learning Approach by Random Forests

The final approach used for keyword extraction is the machine learning classification task which intuitively modeled with the findings in the previous steps. For this classification task we define features which illustrates a keyword candidate [25]. Again, keyword candidates are constructed by forming 1, 2 and 3-gram word tokens. Features are described in table VI. Here each tourist review of a restaurant is considered as a single document and again *Arendsnest Dutch Beer Bar* dataset is used to make computations. As it is discussed in the beginning of the paper, each tourist review (document) has a title and a review text.

Since this is a supervised classification task the keywords for this sample is picked by volunteers and later used in the labeling stage. Categorical data is dealt by hot-encoding in order to be ready for training model. After necessary manipulation in data frame, it is decided that the keyword occurrence is a very rare event with a class ratio of 7:10000. Over sampling and under sampling may be applied in order to overcome the imbalance. Besides, decision tree classifiers tend to give promising results by generating rule based algorithms. Next, the data is split having 2000 observations of test data and around 9000 training data. They have 5 and 3 keywords in their samples respectively.

CART Decision Tree Algorithm is used in training. Testing the model gave 100% results in all the following metrics: accuracy, precision, recall and F1 score. All the keywords are predicted correctly.

TABLE VI
FEATURES OF A KEYWORD CANDIDATE

	Explanation	Type
Name	Name of the candidate	String
TF (Term Frequency)	Total count of the candidate / Max count in Document	Numerical
IDF (Inverse Document Freq.)	# of documents having the candidate/ total # of dociments	Numerical
TOR (Title Occurrence Ratio)	# of titles having the candidate/ total # of titles	Numerical
ROR (Review Occurrence Ratio)	# of reviews having the candidate/ total # of reviews	Numerical
POSS (Part-Of-Speech Sequence)	Part-of Speech Sequence Tag ex. {NN}	Categorical
Ngram (N-gram Tag)	Unigram, Bigram or Trigram	Categorical
Keyword (Target Value 1-0)	if the candidate is a keyword 1; else 0	Binary

VIII. RESULTS & DISCUSSIONS

All of the approaches in keyword extraction part has their own advantages and disadvantages. RAKE is very fast and easy but does not go beyond picking frequent words and does not take other features into account. Key-Graph is a strong visualizer which allowed us to see relationships between words but its expertise is not a primary concern for case since restaurants are described by relatively shorter independent sequence of words. Our final approach, machine learning by decision trees is the most promising one due to its high performance scores. In contrast, it is cost expensive in the terms of computing features of large candidate set and labeling data.

The flow of the proposed system:

1. Online data is collected from multiple sources.
2. Each entity; restaurants and touristic events are scored based on sentiment scoring.
3. Each entity's keywords are generated by the machine leaning approach. Later all of these keywords are gathered to form a text cloud.
4. Users will be asked to pick n desired keywords from that text cloud. These keywords will be the core inputs of the system, additional inputs such as desired money to be spent or the hourly time range that the user would like to be spending can also be added.
5. By taking primarily the keywords input and additionally other extra inputs, the system will generate an optimal automated day plan by using the sentiment scores that are stored and constraints that are defined by the user.
6. The system will also output the overall satisfaction score and the average estimated cost of that plan.
7. Users may discard an entity on the recommended day plan with an option of with or without replacement. They can even discard the whole optimally recommended day plan to go with the next optimal one.

IX. CONCLUSION

Among used methods Naïve Bayes gave satisfied results with a balanced training data set, whereas TF-IDF approach failed to perform well at predicting negative reviews. Later three different approaches are tested on a sample restaurant data for keyword extraction. The extracted keywords will be the descriptive tags of each restaurant and touristic event, hence it plays an important role in recommender system development. It should be noted that each and every step of this study can be applied to any touristic destination as long as there is available data online. Additionally, the proposed recommender system will be developed to combine a variety of review sources.

X. REFERENCES

[1] TripAdvisor, retrieved from https://www.tripadvisor.com.tr/Tourism-g188590- Amsterdam_North_Holland_Province-Vacations.html (Last access: 22.05.2016)

[2] Pang, B., Lee, L., Vaithyanathan, S. Thumbs up?: Sentiment Classification Using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 79-86, July 2002.

[3] Metsis, V., Androutsopoulos, I., & Paliouras, G. Spam filtering with naive Bayes – Which naive Bayes? Third Conference on Email and Anti-Spam (CEAS), 2006.

[4] Schneider, K.M. On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification, in Proceedings of the 4th International Conference on Advances in Natural Language Processing, Alicante, Spain, October 2004, 474-485.

[5] Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R. Tackling the Poor Assumptions of Naive Bayes Text Classifiers, Proceedings of the 20th ThInternational Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[6] Stanford Natural Language Processing on Coursera: https://www.coursera.org/course/nlp (Last access: 22.05.2016)

[7] Raschka, Sebastian. (2015) Pyhton Machine Learning. Birmingham, UK: Packt Publishing

[8] Paltoglou , G., Thelwall, M. A study of information retrieval weighting schemes for sentiment analysis, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 1386-1395, July 2010, Uppsala, Sweden

[9] Ghag, K. and Shah, K. (2014). SentiTFIDF – Sentiment Classification using Relative Term Frequency Inverse Document Frequency. (IJACSA) International Journal of Advanced Computer Science and Applications

[10] Performance Measures for Machine Learning, retrieved from http://www.cs.cornell.edu/courses/cs578/2003fa/performance_measures.pdf (Last access: 22.05.2016)

[11] Kimono + MonkeyLearn: sentiment analysis with machine learning and web scraped data, retrieved from https://blog.monkeylearn.com/kimono-monkeylearn-sentiment-analysis-with-machine-learning-and-web-scraped-data/ (Last access: 22.05.2016)

[12] Witten, H., Ian. Text Mining. Computer Science, University of Waikato, Hamilton, New Zealand.

[13] Bansod, R., Mangrulkar, R. & Bhujade,, G. Text and Image based Spam Email Classification using an ANN Model- an Approach. International Journal on Recent and Innovation Trends in Computing and Communication.

[14] Part-of-Speech Tagging [PowerPoint Slides]. Retrieved from https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf (Last access: 22.05.2016)

[15] Part-of-Speech Tagging [PowerPoint Slides]. Retrieved from http://www.computational-logic.org/iccl/master/lectures/summer06/nlp/part-of-speech-tagging.pdf (Last access: 22.05.2016)

[16] Brants, Thorsten. TnT A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP -2000, April 29 – May 3, 2000, Seattle, WA.

[17] Ritter, A and et al. Named Entity Recognition in Tweets: An Experimental Study. Computer Science and Engineering University of Washington Seattle, WA 98125, USA.

[18] Introduction to Sentiment Analysis [PowerPoint Slides]. Retrieved from http://lct-master.org/files/MullenSentimentCourseSlides.pdf (Last access: 22.05.2016)

[19] Clark., J., H. and Gonzales-Brenes, J., P. Coreference Resolution: Current Trends and Future Directions. November 24, 2008.

[20] Searle, J., R. (2010) Making The Social World: The Structure of Human Civilization. New York, NY: Oxford University Press.

[21] Word Sense Disambiguation. Retrieved from http://www.scholarpedia.org/article/Word_sense_disambiguation (Last access: 22.05.2016)

[22] Tripathi, S. and Sarkhel, J., K. Approaches to Machine Translation. Annals of Library and Information Studies Vol. 57, December 2010, pp388-393.

[23] Information Extraction. Retrieved from https://en.wikipedia.org/wiki/Information_extraction (Last access: 22.05.2016)

[24] Ohsawa, Y., Benson, N., E. & Yachida, M. KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. Graduate School of Engineering Science Osaka University, Toyonaka, Osaka 560-8531, Japan.

[25] Chen, C. Using Random Forest to Learn Imbalanced Data. Department of Statistics, UC Berkeley.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 49–56
DOI: 10.15439/2016F211
ACSIS, Vol. 9. ISSN 2300-5963

A practical study of neural network-based image classification model trained with transfer learning method

Marek Dąbrowski, Justyna Gromada, Tomasz Michalik
Orange Polska, Centrum Badawczo-Rozwojowe, ul. Obrzeżna 7, 02-691 Warszawa
Email: {marek.dabrowski, justyna.gromada, tomasz.michalik}@orange.com

Abstract—This paper deals with algorithms for image classification, which aim to guess "what is on the picture" using human-readable labels or categories. A supervised learning approach with Convolutional Neural Networks (CNNs) is studied as an effective solution to different computer vision problems, including image classification. Main contribution of this paper is a set of practical guidelines to tackle the image classification problem using publicly available tools and typical hardware platforms.

I. Introduction

MACHINE learning is an area of computer science, which assumes that a computer program may "learn" by experience. In other words, the performance of a program to do its job may improve thanks to observation and analysis of actual data samples. As an example, carefully crafted artificial neural network may learn to classify images (name an object on a photo), after it has been trained with sufficiently large number of labeled image examples.

A. Computer vision

The ultimate goal of computer vision research is to teach computers to see and understand images, in a similar way as humans do. Remark that normally the images (photographs, sketches, figures) are represented in computer memory as sets of pixels, and more precisely as bytes with value depending on color intensity of particular pixel. The computer representation of an image does not normally hold any semantic information, unless the user has explicitly provided some semantic context metadata. Thus, it is very difficult for traditional computer algorithm to guess what is the semantic content of a picture.

Machine learning approach has recently been successfully used for solving numerous problems related with understanding of images:

- **Image classification**. A basic problem of computer vision, with goal of characterizing given image by assigning it a human-understandable text label (what is on the picture – is it a person? or a dog? or a building?). The label is usually chosen from a known set of categories, thus this problem is of "classification" type.
- **Classification with localization**. Instead of assigning a single label as in basic image classification, multiple labels may be more appropriate if an image displays not one, but several objects. Additionally, the localization function provides coordinates (bounding boxes) of objects detected on a photo.
- **Object detection**. Quite similar to the previous one, but it rather assumes that we would like to detect presence of a limited set of objects. For example, an object detection algorithm implemented in an autonomous car control system may localize pedestrians (persons) on the side of the road. In this case, there is a single type of object ("person") to be detected.
- **Instance segmentation**. The result of classification with localization is typically a set of identified objects with their approximate bounding boxes (rectangles). Sometimes a more detailed approach would be preferred, being able to assign particular pixels of the image to detected objects. For example, if we detect three persons on a photo, we would have also a precise boundary of each person's shape.

In this paper we focus specifically on the basic problem of image classification.

II. Convolutional Neural Networks: State of the Art

A. Artificial neural networks

Artificial neural network is a biologically-inspired machine-learning model for solving classification problems [1][2]. A simple model of an artificial neuron is presented in Fig.1. A neuron takes a set of inputs (x_i) and produces output value y, applying a non-linear activation function on a weighted sum of inputs. A neural network consists of multiple neurons, connected to each other to form multi-layer structures. If the output is calculated as linear combination of all inputs, such neuron is sometimes called a "fully-connected" unit.

Before being able to actually solve a classification problem, a neural network must be "trained". It means that its parameters (weights) are tuned by a special algorithm, which takes as input a large set of training data. Training data consist of input values together with "ground truth"

result, that is a result, which we know is correct for given input.

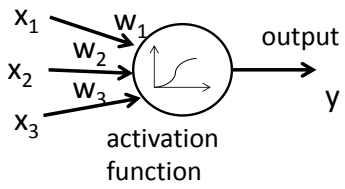


Fig. 1 Model of a simple artificial neuron

A special family of artificial neural networks, called "convolutional", is especially successful for image classification [2]. Let us assume that input to the model consist of certain number of pixels, with value corresponding to color intensity (see Fig.2).

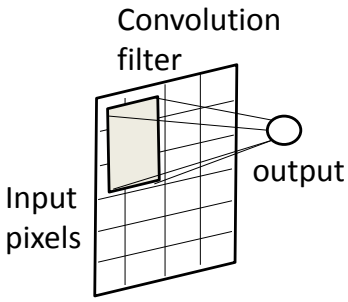


Fig. 2 Model of a convolutional unit

Instead of calculating the output value by taking a combination of all inputs, the output value of convolutional unit is based on a locally constrained region of several pixels (e.g. 3x3, 5x5, 7x7). Moreover, the output is calculated for each local area on the image, like a filter which is slided all over it. The idea is to take into account the structure of computer image and analyze in more detail pixels in a local neighborhood, in order to detect basic shapes, edges and other small characteristic graphical features. Then, a multi-layer hierarchy of convolutional units is applied, being able to detect more high-level shapes and structures of an image.

As a final layer of neural network model, fully-connected neurons are always used. In particular, the last layer consists of k neurons, where k is the number of image classes that the model is able to distinguish. After passing an image (represented as pixels intensity values) through all layers of neural network, the value outputted by k_{th} neuron of the final layer will correspond to estimated probability that the image belongs to the k_{th} class. Thus, this final layer is sometimes called a "classification layer" because it is aware of set of classes of a given model, and is trained to recognize the class of the image based on features that are spotted by internal layers.

B. State-of-the-art models for image classification

The basic artificial neuron structures as depicted on Fig.1. and Fig.2 constitute main building blocks of modern neural networks. But to really appreciate its power for solving classification tasks, we need much more complex structures with thousands, or even millions of such small basic units. Multi-layer structures appear to be more effective, which led to a concept "deep learning", which denotes artificial neural networks consisting of many layers of connected units [1][2].

Looking back into history of research work in this area, a model published in 1998, named "LeNet" [3], is considered as pioneering work in applying convolutional neural networks to computer imaging problems. It had 8 layers (convolutional and fully-connected), with around 1mln parameters in total. LeNet primary application was to recognize handwritten digits in banking information systems.

A new wave of research in neural networks came few years ago thanks to several breakthrough advances: invention of deep networks, more efficient training algorithms, availability of more powerful hardware for computationally intensive calculations, and availability of large data sets for training. In 2012, so-called "AlexNet" model was proposed [14], with 14 layers and about 60mln parameters. In the trend of building deeper networks, the "GoogleNet Inception" model [4] has been proposed in 2014, with 22 layers (Fig.3). It consists of convolutional (blue boxes on Fig.3) and fully-connected units (yellow), together with pooling units (red) and joining units (green). Thanks to more efficient use of convolutional units it had only 5mln parameters, which made it easier to train and run on less powerful hardware, while being more effective than the older models. A newer version of "GoogleNet" architecture will be used in experiments presented later in this paper.

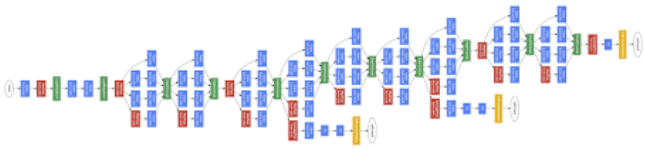


Fig. 3 "GoogleNet": state-of-the-art multi-layer convolutional model for image classification

C. Imagenet image corpus

The neural network models described in previous section could never work well without being trained on a very big sets of learning data. A publicly available ImageNet database [5] is a widely used repository of properly labelled photos (i.e. photos with text label correctly describing its content). The ImageNet corpus contains 21841 classes of images. Each class is described by a label, which is called a "synset" by referring to the WordNet taxonomy of concepts. With an ultimate goal of collecting around 1000 images for each class, the ImageNet repository has now about 14mln human-annotated images.

Once a year the ImageNet team organizes a competition for computer vision researchers to propose world-best models for solving image classification problem. For the purpose of this competition, a subset of 1000 basic ImageNet classes has been selected as a reference corpus. This image corpus with 1000 classes has been used in experiments presented in this paper.

III. TRANSFER LEARNING

The "deep learning" approach adopts artificial neural network models with large number of hidden layers and millions of parameters. For training such models, the backpropagation method seems to work well [2]. Basically, it assumes minimization of cost function (also called loss function), which describes overall difference between the predicted and actual (ground truth) classification results, for a given set of training examples. Several methods are typically used for the optimization algorithm, usually being variations of Stochastic Gradient Descent [2].

The computational cost of the backpropagation method is noticeable, especially that the training sets should be huge, with tens of millions of training examples. Despite of recent advances in hardware processing, distributed computing and Graphical Processing Units (GPUs), which greatly accelerate computations, it may take weeks to train a model until it reaches sufficient accuracy [4].

To avoid the lengthy learning process and allow for training with smaller amount of training examples, so-called Transfer Learning method has been proposed [6]. Transfer Learning is a machine learning method, which is improved thanks to transfer of knowledge from a related task, that has already been learned. In other words, new image classification models may be trained much faster, thanks to using parameters of a previously trained model. The process for practical usage of Transfer Learning is the following (see Fig.4):

1. Get a previously trained model. Pre-trained models published by other researchers are available e.g. from [8][9].
2. Split the old model architecture into two parts:
    a. All hidden layers of the neural network, with their structure (connections) and previously learned weights, will be copied into the new model.
    b. The last layer of neural network, which performs actual classification into one of the classes, is strictly related with the old model and will be disregarded in the new model.
3. Prepare a new set of training examples (images labelled with appropriate class name, as required for the new model).
4. For a new set of training images, calculate the output values after passing through the first part of neural network (the one that is transferred into the new model). The numerical value calculated as output of given image, at the next-to-last layer of original model, will be called a "bottleneck".
5. Add a new final fully-connected layer, which will now constitute the last layer of new neural network model. This new final layer will calculate the probability of given image belonging to a given class.
6. Train the new final layer with previously calculated "bottlenecks" as input, and a set of new "ground truth" labels that denote true classes of training images.

Remark that in this method only the last layer of artificial neural network model has to be trained from scratch, while for all previous layers (and remind that for example the GoogleNet model has 22 of them) the weight values are copied from the previously pre-trained model. Thanks to that, we can create a new image classification model, with our own classes and labels, within several hours instead of weeks, on standard hardware.

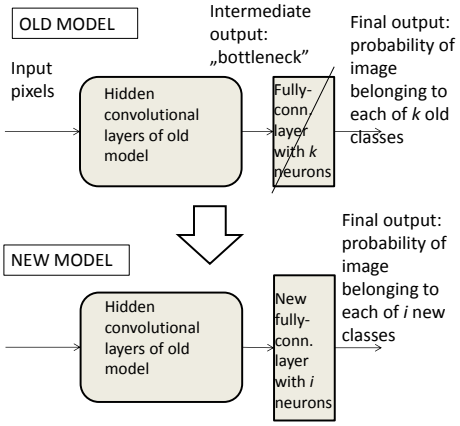


Fig. 4 Illustration of "Transfer Learning" method

IV. DEPLOYMENT SCENARIO

Running Machine Learning algorithms and neural networks especially requires certain amount of computing resources, not only processing power, but also storage and access to big repositories of training data. Thus, a practical deployment approach may assume a networked environment, with computational resources deployed in cloud data center, with Web Services API developed for client applications to upload images and receive results over the web. An example deployment scenario is presented in Fig.5.

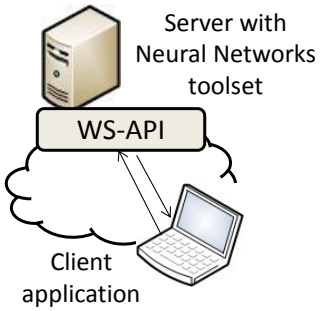


Fig. 5 Networked deployment scenario

V. EXPERIMENTAL SETUP

A. Tensorflow: open source tool for neural networks

For our experiments we have used Tensorflow [9] library for numerical computation. Tensorflow has been created by Google and made available as open source. It supports various types of numerical computations, including complex neural network models, on various types of hardware, CPU and GPU. It supports the Transfer Learning method.

B. Lab infrastructure

It is well known that GPU hardware greatly improves performance of machine learning computations. Having a goal to find optimal hardware setup for our machine learning task (minimize platform cost, assuming satisfactory training and testing time) we have evaluated performance of several typical medium-level hardware platforms.

We wanted to simulate a home or small company environment, in addition to large expensive data center.

1) PC computer: a typical desktop with 4 CPU cores and 8GB RAM. It had a GPU card NVIDIA GeForce GTX 960. Tensorflow computations may or may not use GPU depending on software configuration, so both hardware settings were tested (PC-CPU and PC-GPU).

2) ODROID: „mini" home computer of SBC (Single Board Computer) type [12]. It is a fully-fledged computer with a very small size, low power consumption and low price (see Fig. 6). It is not as powerful as a desktop PC, but has sufficient capabilities for simple home tasks.

- *ARM architecture*
- *SOC: Samsung Exynos*
- *32bit architecture*
- *2GB RAM*
- *Flash disk*
- *USB, Eth, HDMI*
- *Linux*
- *Price:* ~70$

Fig. 6. ODROID: exemplary SBC device (Single Board Computer)

VI. Experimental verification of transfer learning method

The goal of experiments was to verify if the Transfer Learning method, applied to training image classification models, allows for achieving results that are as good as full-fledged training, in much quicker time and with less computational resources. The GoogleNet model [10] trained with ImageNet-1000 corpus has been used as basis for re-training. The ImageNet database of basic 1000 classes has been downloaded and used as re-training examples, following the Transfer Learning concept.

Remark that we have decided to use in re-trained model the same 1000 ImageNet classes as in the original model. At first look it may be counterintuitive: why we would re-train the model to have the same result at the end? Of course, in a target scenario the re-training procedure would assume completely different target set of classes, with different set of training images than the original model. But remind that the goal of experiments presented in this paper was to evaluate the correctness of Transfer Learning and so it seems methodologically correct to compare the re-trained model with the original one, that was created with the same initial assumptions and target set of classes.

A. Testing the image recognition capability

First, let us discuss what is the expected result of image classification. Say, we have a photo, and we would like to tag it automatically with a text label. The result of image classification algorithm should thus be a word, or a few words, matching with certain level of confidence semantic contents of the photo. Strictly speaking, the result of passing the image through neural network model trained with 1000 classes will be a vector of 1000 numbers, corresponding to the "score" associated to each class. The "scores" are conceptually related to a likelihood that given result is a correct one. The distribution of score values for all classes should form a proper probability distribution, so the scores will sum up to 1.0.

An exemplary photo with classification result is presented in Fig.7 (the result has been calculated with one of the models trained in the scope of this study, but this is not important at this moment, as it is presented merely as illustration of intended goal of the algorithm). More precisely, this is a "Top-5" result, that is five class names with the highest values of scores. In our example, the score associated with class "palace" is 0.97, which means that the algorithm thinks with very high confidence that there is a "palace" on the photo, which is actually true.

- "palace" (0.97)
- "monastery" (0.11)
- "fountain" (0.0036)
- "castle" (0.0022)
- "church, church building" (0.0019)

Fig. 7. Result of re-trained classification algorithm on a real-life photo example

This simple example shows that image classification algorithm put in the realistic setting produces quite accurate results. However, we need more rigorous and repetitive method to evaluate objectively the correctness of the method, as applied to a larger set of images.

Assume that we have a test set of N images, drawn from the ImageNet corpus and thus human-annotated in controlled way with a "ground truth" label. Remark that according to Machine Learning established practice, the examples from the test set must not be previously used as training data, since that would bias the test result towards positive outcome. For all images of the test set, the test result is produced by the evaluated algorithm, in the form of 5 classes with maximum associated scores among the all the k classes.

The metric "Top-1 accuracy" will be defined as ratio of correct classification results in the entire test set (where positive result means: "the label with maximum score is equal to ground truth"). The Top-1 accuracy metric may be too restrictive in realistic scenarios. Take an example in Fig.1 – "palace" is a true object on the image indeed, but if you just look at the photo, without prior knowledge of what it actually is, it may appear that "castle" could also a possible name for an object on photograph. Thus, another metric, "Top-5 accuracy" has been defined to take into account that sometimes the ground truth label may be among the best recognized, but not necessarily at the first place. Top-5 accuracy is also defined as ratio of correct classification results in the test set, but the positive result is now: "the ground truth label is among the 5 maximum-score labels assigned by the classification algorithm".

We have validated the studied model re-trained with Transfer Learning method on a test set of 50000 random images from the ImageNet corpus. The obtained Top-1 and Top-5 accuracy metric results are presented in Table 1.

TABLE I. RESULT OF FINAL EVALUATION OF RE-TRAINED MODEL ON A TEST SAMPLE OF 50000 IMAGES FROM IMAGENET CORPUS

Metric	Test result
Top-1 accuracy	88.2 %
Top-5 accuracy	98.0 %

These results for re-trained model are impressively good. In fact, for 88% of ImageNet photos the algorithm is able to classify correctly the true label, while for 98% the true label is among top 5 assigned ones. Surprisingly, the achieved accuracy appears even better than the accuracy of original model, reported in [10] (top-5 accuracy equal to 93.33%). We think that the reason for this misleading result is that for testing the re-trained model we have used the ImageNet images that belonged to the training set of original model. Thus, the test result is higher than expected, because in some sense the artificial neural network is tested with images that it has already seen before. Unfortunately, the test set that was used by authors of the original model in [10] is not publicly available, so we were not able to make a truly relevant comparison. Nevertheless, we think that the conclusion that can be drawn from our study is that re-training the artificial neural network using Transfer Learning method may give us a model that is as good as the original one, in relatively short time on typical modern computers (see later in the paper).

B. Finding a parameter setup for transfer learning method

There are several parameters that we can shuffle in order to obtain a satisfying model within reasonable training time. Our goal was to achieve well trained model (not over fitted, nor under fitted) with the test accuracy around 90% for Top-1 accuracy metric and close to 100% for Top-5 accuracy metric. The intended time for re-training a model on our infrastructure was max 12h, which meant running experiments taking 32000 training steps which is around 3 epochs for our training data set.

The main parameters that we tuned were: type of the optimizer and value of learning rate. The optimizer takes the loss computed in forward propagation part (in our case, it is the loss function for softmax classifier [1], called cross-entropy loss), calculates the gradients in backward propagation and then changes the weights of the model trying to minimize the loss. In case of Transfer Learning algorithm, the optimization process concerns only one layer – the last one. The learning rate value is a hyper-parameter which tells how fast the optimizer should converge to minimal loss. When the learning rate is to low, the process of training may last very long to achieve optimal values or never achieve it but on the other hand when the learning rate is too high, the average loss may increase which is opposite to our goal. Hence, the choice of learning rate and appropriate optimizer is crucial.

We have chosen two optimizers for our experiments: Stochastic Gradient Descent Optimizer (SGD) [13] and Adam Optimizer (Adam) [11] and used them in training process with various values of hyper-parameter learning rate $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. To choose the best configuration of these parameters, we were observing how the cross-entropy loss, training Top-1 accuracy and validation Top-1 accuracy behave during the training process (Fig.8-11).

When we look on the graph of cross-entropy loss in function of training steps (Fig.8), three potential candidate configurations seem the most promising: the Stochastic Gradient Descent Optimizer with learning rate $\alpha = 0.1$, Adam Optimizer with $\alpha = 0.05$ and Adam Optimizer with $\alpha = 0.01$. The cross-entropy loss in these cases constantly goes down and within 3 epochs reaches the value of around 0.36. The cross-entropy loss for Stochastic Gradient Descent Optimizer with learning rate $\alpha = 0.01$ also goes down but much slower than in previous three configurations. The cross-entropy loss for Adam Optimizer with learning rate $\alpha = 0.1$ after 2000 training steps goes up, which is undesirable behavior and may suggest too large learning rate.

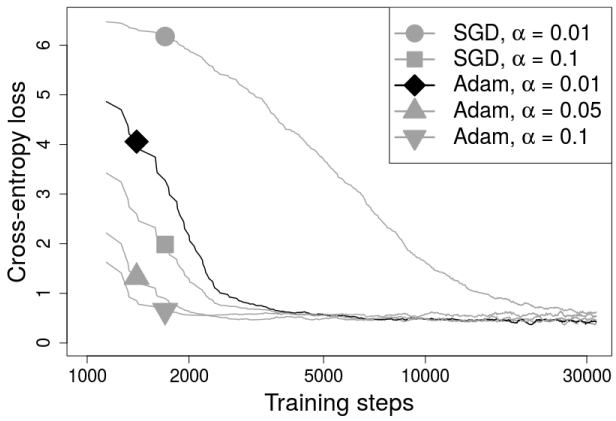


Fig. 8. Entropy in function of training steps (log scale) for Stochastic Gradient Descent Optimizer (SGD), Adam Optimizer (Adam) and various learning rates (α)

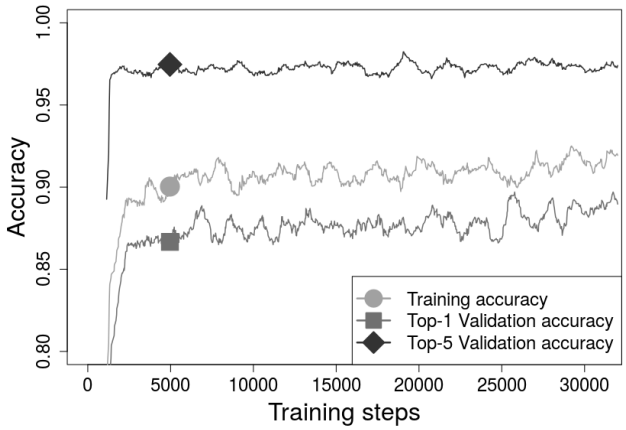


Fig. 9. Accuracy for Stochastic Gradient Descent Optimizer and learning rate α = 0.1

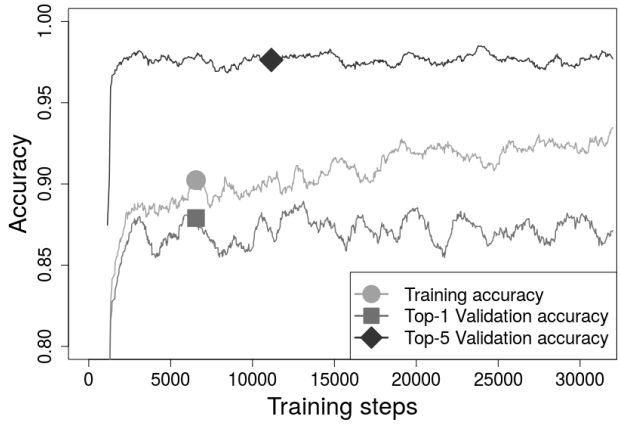


Fig. 10. Accuracy for Adam Optimizer and learning rate α = 0.05

The Top-5 Validation accuracy for all three chosen configurations achieve very quickly a desirable value of around 97% so we will focus on training accuracy and Top-1 validation accuracy when comparing the pointed solutions.

For Stochastic Gradient Descent Optimizer with learning rate α = 0.1 (Fig.9), we can observe a big gap between the training accuracy and Top-1 validation accuracy.

For Adam Optimizer with learning rate α = 0.05 (Fig. 10), the situation is a bit worse: the gap between training accuracy and Top-1 validation accuracy is not only big but it is even increasing.

The big gaps between training accuracy and Top-1 Validation accuracy for both configurations: Stochastic Gradient Descent Optimizer with learning rate α = 0.1 and Adam Optimizer with learning rate α = 0.05 suggest that the models might be over fitted.

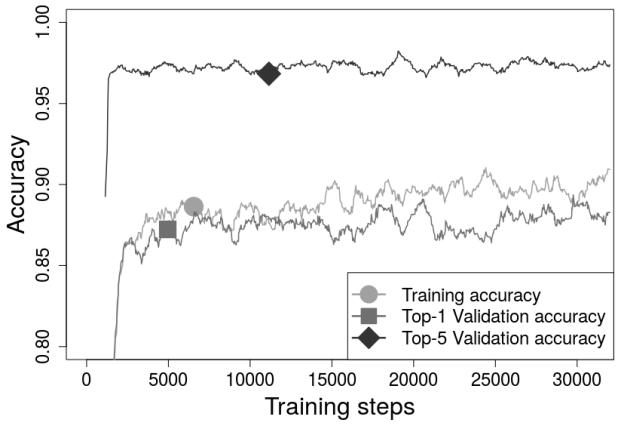


Fig. 11. Accuracy for Adam Optimizer and learning rate α = 0.01

The last candidate: Adam Optimizer with learning rate α = 0.01 (Fig.11) gives both acceptable values for Top-1 validation accuracy of around 0.87 and acceptable gap between the training accuracy and Top-1 validation accuracy. What is more, the figure of cross-entropy loss in this case also seem the best: it reaches the lowest values after 10000 training steps and constantly goes down. Hence, this configuration has been chosen as a final setup for Transfer Learning method.

Summarizing, chosen configuration of parameters for Transfer Learning method is as follows: training steps= 10000, Adam Optimizer with learning rate α=0.01 and epsilon=0.1 (a small constant for numerical stability), train batch size=100.

VII. PERFORMANCE BENCHMARKING

Next, we have performed a series of experiments to benchmark hardware configurations available in our lab (see section V B) as platforms for re-training image classifier models. As discussed in section III, we can distinguish several phases in the process of Transfer Learning method and so the performance benchmarks were performed separately for each one of them.

A. "Bottleneck" pre-calculation phase

First computationally intensive step is to determine the "bottleneck" values, that is an output of pre-last layer of neural network, given a single image at the input. Measured time of performing that operation, averaged over 50 test runs for different images, is presented on Fig.12. Not surprisingly, the Odroid platform is the slowest one, with 2s to calculate a single bottleneck. The PC platform with GPU card is a clear winner, with 0.08s time. Remark that the bottleneck calculation time includes the time to read the image file from disk. On all studied platforms this time was below 0.004s, so we consider it negligible.

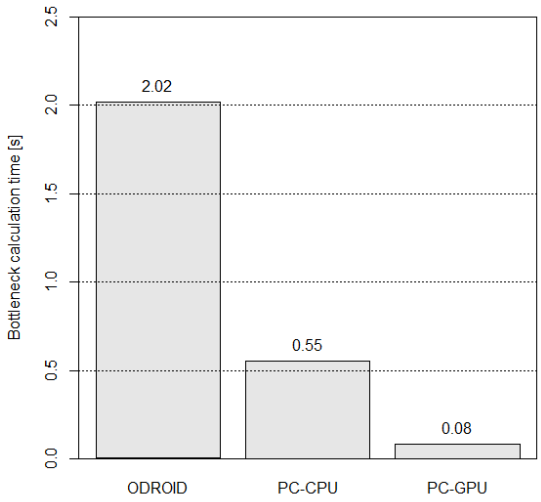


Fig. 12. Calculation time of "bottleneck" value for 1 image

The bottleneck calculation operation has to be done once for every image in the training corpus. The result can be saved in a file, for quick access during the re-training phase later. Taking that into account, Table II presents estimated total time to calculate bottlenecks for all images in the training set, for different number of target classes, and assuming approximately 1000 images per class. We can see that for a 1000-class model, it takes about 1 day on our fastest platform to prepare "bottleneck" values for entire image corpus. Fortunately, this operation is done just once, with results saved on disk for quick access during re-training phase.

TABLE II. ESTIMATED BOTTLENECK CALCULATION TIME FOR WHOLE IMAGE CORPUS

Number of classes	ODROID	PC-CPU	PC-GPU
100	56h	15.3h	2.2h
250	140h	38.3h	5.7h
1000	23d 8h	6d 9h	22.7h

B. Re-training phase

Now, the final classification layer of neural network is being trained with training images. The process follows a Stochastic Gradient Descent algorithm [2] with "mini-batch". In each step, the values of loss function, gradients and weight updates are calculated for a batch of images, in our tests equal to 100, 1000 and 10000. Measured average time of training step, normalized for 1 training image (i.e. divided by batch size) is presented in Fig.13. We can distinguish the following time components:

- **File access time** is the time to read the "bottleneck" value from an a-priori saved file, as discussed in section VIIA. Remark that this time component is highly related with performance of disk file system, rather than processor computing power. It is a little bit surprising that it has the biggest impact on total training time. Remark that at each training step, a large number of small files (equal to batch size: 100, 1000, 10000) is read from disk and then processed in-memory. We may presume that optimization of disk access, e.g. by using faster disks, or pre-caching all training data in memory could bring significant reduction of this phase (a single "bottleneck" file has about 18kB size, which means that for entire image corpus we need about 18GB - unfortunately for current experiments we didn't have a machine with sufficient amount of RAM).
  Still, we can see that the time to access pre-cached bottleneck data is still much smaller than the time to produce the data by doing full calculations, as measured previously (about 0.015s to read from disk, vs. 0.08s of calculations on fastest machine). This result confirms that it indeed makes sense to pre-calculate and save "bottleneck" data for later usage multiple times in re-training.
- **Training time** is the actual time to perform all mathematical calculations related with completing the training steps. This time appears to be negligible comparing with the time of accessing training data from disk.
- **Validation time**. Once every 500 batches, the validation step is performed, i.e. the value of top-1 and top-5 accuracy metric is calculated for a given validation set. This operation is done for the purpose of monitoring and logging the learning process, and has no impact on final result. But since it is usually done, we report it also in time benchmarks.

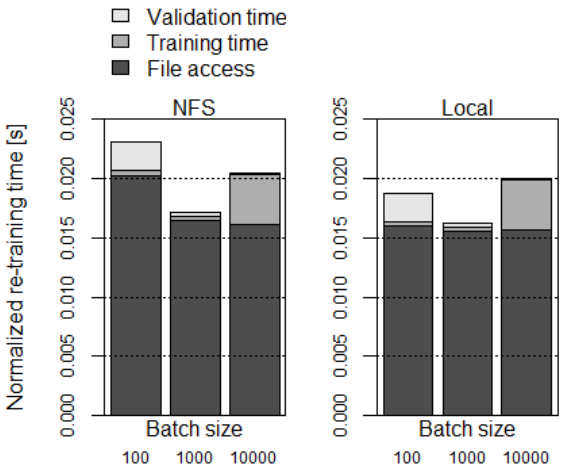


Fig. 13 Analysis of re-training time, normalized for 1 training example

The results presented in Fig.13 correspond to processing a single training image. For completeness, Table III presents measured total time of re-training with a corpus of images appropriate for a given number of classes (recall that we have around 1000 images per class). The reported time corresponds to 1 "epoch", that is training once with a full set of training data. Due to the fact that file access time has greatest impact on training time in this phase, the differences

between hardware platforms are not so big. Surprisingly, the less powerful device (Odroid) performs best for this task. The explanation is that this device has a fast flash disk which beats in terms of file access time (which, as we saw has great impact on performance) the magnetic disc of our lab PC.

TABLE III. MEASURED TRAINING TIME OF 1 EPOCH (AVERAGE AND STANDARD DEVIATION)

Number of classes	ODROID	PC-CPU	PC-GPU
100	8.57 ± 0.57 m	14.24 ± 0.08m	14.57 ± 0.57m
250	22.91 ± 0.69m	36.55m ± 1.58m	36.66 ± 1.53m
1000	2h7m ± 3.18m	3h39m ± 16.35m	3h30m ± 8.14m

C. Test phase

Finally, test phase is when we want to actually obtain a classification result for an arbitrary photo. Fig.13 presents averaged measurement of time to calculate a final result. The following phases are distinguished:

- **Session run time**: actual time of running the calculations with image pixel values as input, and scores assigned to each class (there were 1000 classes in tested model), as output.
- **Result processing time**: time needed to prepare the result, i.e. sort the result table to extract 5 best scores, lookup the labels table to retrieve human-readable names of classes, and prepare the final result as json structure.

As expected, total test time is longest on the Odroid (session run time is around 6s). On the other hand, on a PC with GPU this processing is time is reduced to less than 2s.

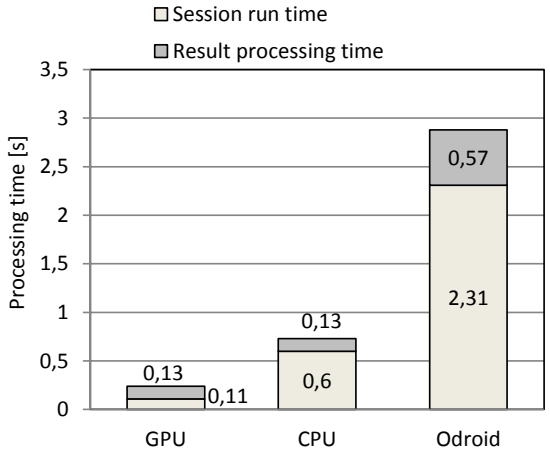


Fig. 14 Analysis of inference time (image testing) on different hardware platforms

VIII. SUMMARY AND FUTURE WORK

Transfer Learning method has been studied as efficient approach to training neural network models for image classification. Practical guidelines for setting configuration parameters of re-training process were given in the paper, which includes: number of training steps for achieving sufficient accuracy of re-trained model, type of optimization algorithm, and value of learning rate. Performance benchmarks for training image classification models on typical low- and middle-level hardware platforms were given. The measurements show significant advantage of using GPU card for computationally demanding operations, and, a little surprisingly, advantage of low capacity device equipped with ultra-fast disk, in the case of training phases where lots of training data must be accessed in short time. In future work we plan to extend the benchmarking experiments to cover wider range of hardware platforms, including machines with different CPUs, RAM size, and different GPU types.

It was shown that a good quality image classification model with our own set of classes can be obtained in several hours instead of weeks, by applying Transfer Learning method, that is re-training an existing neural network model downloaded from publicly available source and re-using most of the parameter values of original model. In future work we plan to study transfer learning with different sets of images than basic ImageNet-1000 corpus, with a goal to improve realism of our scenarios and transferring learned features to completely different classification task.

REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "Deep Learning", Book in preparation for MIT Press, 2016, on-line version available at:http://www.deeplearningbook.org
[2] Michael A.Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015, on-line version of the book available at: http://neuralnetworksanddeeplearning.com/index.html
[3] LeCun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., Henderson, D.,Howard, R. E., and Hubbard, W.. Handwritten digit recognition: Applications of neural network chips and automatic learning. IEEE Communications Magazine, 27(11), 1989
[4] Ch.Szegedy et al, "Going deeper with convolutions", http://arxiv.org/abs/1409.4842
[5] ImageNet database of computer images: http://image-net.org/
[6] Yosinski J, Clune J, Bengio Y, and Lipson H. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems 27 (NIPS '14), NIPS Foundation, 2014
[7] Caffe Model Zoo web page: https://github.com/BVLC/caffe/wiki/Model-Zoo
[8] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo,
[9] Z. Chen, et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
[10] Ch.Szegedy et al., "Rethinking the Inception Architecture for Computer Vision", http://arxiv.org/abs/1512.00567
[11] D.Kingma, J.Ba, "Adam: A Method for Stochastic Optimization", http://arxiv.org/abs/1412.6980
[12] ODROID-XU4 hardware : http://www.hardkernel.com/main/products/prdt_info.php?g_code=G143452239825
[13] Y. LeCun, L. Bottou, G. Orr and K. Muller: Efficient BackProp, in Orr, G. and Muller K. (Eds), Neural Networks: Tricks of the trade, Springer, 1998
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In NIPS 2012, Neural Information Processing Systems, Nevada, 2012

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 57–61
DOI: 10.15439/2016F579
ACSIS, Vol. 9. ISSN 2300-5963

Machine Vision in Food Recognition: Attempts to Enhance CBVIR Tools

Andrzej Śluzek

Khalifa University, Abu Dhabi Campus
P.O. Box 127788, Abu Dhabi, UAE
Email: andrzej.sluzek@kustar.ac.ae

***Abstract*—Visual identification of complex images (e.g. images of food) remains a challenging problem. In particular, *content-based visual information retrieval* (CBVIR) methods, which seem a natural choice for such tasks, are often constrained by specific characteristics of the images of interest and (possibly) other practical requirements. In this paper, a novel CBVIR approach to automatic food identification is proposed, taking into account characteristics of solutions currently existing in this area. Based on limitations of those solutions, we present a scheme in which a co-occurrence of MSER features extracted from three color channels is employed to build a *bag-of-words* histogram. Subsequently, food images are matched by detecting similarities between those histograms. Preliminary tests on a recently published benchmark dataset UNICT-FD889 reveal certain advantages of the scheme and highlight its limitations. In particular, a need of a novel methodology for segmentation of food images has been identified.**

I. Introduction

Unstoppable (and sometimes excessive) presence of mobile devices in everyday activities is understandably followed by development of IT tools which can meaningfully analyze and interpret data collected during those activities. The analysis of visual data is particularly important because such data can be easily an unobtrusively captured (almost) everywhere and (almost) continuously in large quantities. Therefore, there is a growing interest in development of applications for the analysis of diversified categories of everyday-life images and videos. Not surprising, one of attractive and prospectively popular categories is food.

Although it cannot be claimed that automatic recognition of food images/photos becomes a research area of very high importance, growing numbers of publications on this topic can be noticed, e.g. [1], [2], [3], [4], [5], [6], which confirms a certain level of interest.

Some of the presented results (e.g. [2], [7], [8]) target specific health-related applications, i.e. diet assessment, preventing obesity, monitoring food allergies, etc., while others (e.g. [3], [4], [5], [6]) just evaluate applicability of diversified machine vision techniques and algorithms to this particular area.

There are also some works proposing benchmark datasets of food images to evaluate various approaches on those datasets and to stimulate research in the area (e.g. [1], [9].

With the universal availability of smartphones (and more advanced wearable cameras expected in the near future) other practical applications of automatic visual food recognition and identification can prospectively emerge, including quality assessment (serving quality or conformity with presentation standards), search for restaurants serving previously seen dishes, etc.

In this paper, we first (in Section II) briefly overview the most typical approaches to visual food recognition, appraise their advantages, and highlight limitations. In particular, we focus on techniques exploiting the most typical mechanisms of *content-based visual information retrieval* (CBVIR), i.e. local feature (keypoint) detection, description and matching, combined (sometimes)with other algorithms. However, the reported results cannot be considered fully satisfactory. Thus, we attempt to investigate an improved low-level mechanism to enhance reliability of such systems. In Section III, a CBVIR-based approach employing neighborhood dependencies between keypoints extracted from three channels of color images is proposed and preliminarily verified. Unfortunately, the conclusions obtained from the experimental results are not too encouraging either. It seems that food recognition based only on currently existing machine vision techniques cannot reach the level of performances already achieved in other areas of visual data analysis. Apparently, identification of food items using only vision is not as straightforward as expected; some suggestions regarding the future works are discussed in the final Section IV.

II. Vision-based Techniques for Food Recognition

The earliest attempts to identify food items from their pictorial representations were rather restricted to relatively narrow categories of food represented by individual items, and sometimes discussed from the robotic perspective. For example, a survey of techniques for detecting individual fruits on complicated backgrounds (for machines automatically harvesting fruits) was discussed in [10]. Another example of such a system, i.e. a vision systems for the identification of broken biscuits on a production line was presented in [11].

Those early techniques are typically based on a preliminary detection of predefined shapes (mostly circles or ellipses) followed by the analysis of color and/or texture properties within those extracted shapes. In many cases, the usage of a supplementary range sensors was additionally assumed (driven by the robotic needs). Usually, the realistic scenarios of

natural conditions were taken into account, including shadows, unusually bright areas, and object overlapping.

In the following years, more attention was paid to algorithms based on local features and their descriptors (instead of just shape and texture/color characteristics) sometimes supplemented by other, more or less sophisticated image processing tools. One of the best known works based on local features was presented in [9]. The authors combined three algorithms. First, SIFT keypoints, [12], were detected and converted into visual words. Then the whole image was represented by a *bag of visual words* (BoW) histogram, [13], and by another histogram of color distribution. Nevertheless, reliability of this approach was rather unsatisfactory. Even though only a few major ingredients of various fast-food items were considered, accuracy of recognition was below 20% on the dataset of fast-food images defined in the same paper.

The works on the identification of fast-foods were continued in [5] and [6]. In both cases, the same dataset of fast-food images were used. However, the authors of [5] rejected the concept of using SIFT local features or color histograms. Instead, they focused on features characterizing local texture properties of images. The features were build over pairs of pixels (with the feature significance inversely proportional to the distance between pixels). Then, a histogram of pairwise features was the major tool for food detection of eight basic fast-food components (e.g. cheese, bread, egg or beef). Based on detection of those components (and their relative localizations) images were classified into 61 categories, for which the reported accuracy reached 28.2% (for the most successful algorithm of relative localization which supported the basic component identification). When only identification of the seven broader categories (e.g. sandwiches or salads) was considered, the accuracy varied between 69% and 78%.

In [6], the concept of using keypoints and the corresponding visual words was revived. However, instead of SIFT keypoints *local binary patterns* (LBP, [14]), which were considered more suitable for texture characterization than SIFTs, were used. The accuracy for the seven broad categories of fast-foods varied from 56% to 90%. The authors also found that for some categories performances of the method based on SIFT keypoints were superior to LBP.

Altogether, even for a narrow category of fast-food items only there is no clear picture regarding the recommended techniques for visual identification of food. The situation is more complicated if a wider range of food items is to be considered. For example, in [4], 85 diversified items of Japanese food were analyzed. The author used a fusion of several image features, i.e. SIFT-based BoWs, Gabor features and color histograms, and applied multiple-kernel learning techniques to achieve accuracy exceeding 60%.

In [3], a wider dataset of 100 Japanese dishes was considered, in which each dish may contain two or more food items (e.g. fish and chips with salad). The food images were preliminarily segmented into individual items using trained classifiers to detect and evaluate segmentation regions, and each detected region was recognized by a multiple-kernel learning technique. Again, usefulness of SIFT features and color histograms was acknowledged there, but they were combined with HoG and (similarly to [4], Gabor texture features.

Currently, it seems the most comprehensive (in terms of the number of food items) study was presented in [1]. UNICT-FD889 dataset of almost 900 diversified dishes (see examples in Fig. 1) have been collected and used to compare and benchmark performances of food recognition algorithms based on various representation models.

Fig. 1. Exemplary images of UNICT-FD889 dataset (from [1])

The major conclusion was that CBVIR approaches (i.e. methods based on keypoint-like feature detection and image matching) are, in general, applicable to food images. Three types of feature-based representations were eventually selected, namely SIFT, Bag of Textons ([15]) and pairwise rotation invariant co-occurrence linear binary patterns ([16]).

Bag of Textons was preliminarily found superior, but the other two representations were not far behind (especially if applied in the variants intended for color images). Nevertheless, the authors did not apply any trained classifier. Therefore, the results were generally inferior to those reported in the previously discussed papers. However, such an approach seems more practical since it would be impossible to have a classifier for each newly encountered food category (e.g. dishes seen the first time).

III. Matching Food Images

A. Methodologigal principles

In this paper we propose a novel method for matching food images, that is conceptually more similar to [1] rather than to the other papers discussed in Section II. Therefore, no classifiers have been built for the known food categories, and the method is open to unknown types of food without any modification or retraining . In general, we assume that:

1) Images are represented by collections of local features which are subsequently converted (through quantization of their descriptors into *visual words*) to *bags of words* (BoW) histograms.

2) The features represent images in a wider visual context, i.e. feature descriptors characterize features in conjunction with a number of neighboring features. Moreover, the features are separately extracted from individual color channels so that those semi-local characteristics incorporate the color properties as well.
3) At the current level of development, images are compared using only BoW histograms, but the future developments may incorporate more sophisticated usage of the features and their descriptors.

Actually, the selected features are MSER keypoints (see [17]) which have low complexity and good performances. They are usually represented by elliptic approximations so that features of elongated shapes (which are frequently present in images of food) can be more accurately handled (compared, for example, to shapes represented by SIFT keypoints).

MSER features are separately found in three color channels (i.e. R, G and B images) and, subsequently, they are analyzed semi-locally using the method similar to the approach preliminarily outlined in [18] and applied in a more sophisticated version in a number of later works (e.g. [19], [20]).

Thus, for any MSER feature of *C* color (where *C* = R, G or B) represented by E_C ellipse we define its *S*-color neighborhood as a collection of *S*-color MSERs (where *S* = R, G or B, and $S \neq C$) as follows:

An *S*-color MSER represented by E_S ellipse belongs to *S*-color neighborhood of E_C ellipse if:

1) The distance $d(E_C, E_S)$ between origins of E_C between K and E_S satisfies:

$$1/2 \times r_{norm} \leq d(E_C, E_S) \leq 2 \times r_{norm}, \qquad (1)$$

where $r_{norm} = \sqrt{area(E_C)/\pi}$.
2) The areas of E_C and E_S ellipses are similar but E_C is larger (i.e. the ratio is between 0.5 and 1).

Using a large collection of images (including a significant percentage of food images) we have found that the average size of such neighborhoods is between 8 and 10.

In each color channel, individual MSER features (i.e. their ellipses) are represented by SIFT descriptors in RootSIFT variant which has been found superior (see [21]). Subsequently, RootSIFT descriptors of the keypoint ellipses E are quantized into visual words $w(E)$ from a vocabulary of either 128 or 1024 (two variants have been implemented) words.

Then, for each keypoint with E_C ellipse we take into account its two *S*-color neighborhoods (containing a number of E_S ellipses). For example, if $C = R$ the neighborhoods will be built using MSER keypoints from *green* and *blue* channels.

Eventually, pairs of visual words $w(E_C)$ and $w(E_S)$ are formed, and each such a pair is described by a word from a vocabulary of either $128 \times 128 = 64k$ or $1024 \times 1024 = 1M$ words. Thus (because the average size of neighborhoods is between 8 and 10) each keypoint of *C*-color contributes, in average, $16 - 20$ visual words to the *bag-of-words* (BoW) histogram of the image.

Finally, similarities between images are estimated by the similarities between their BoW histograms built according to the above principles.

Because our approach is not restricted to images of known and predictable food items, BoW normalization techniques requiring database statistics (e.g. *td-idf*, [13]) cannot be applied, and we use histograms of *absolute* word frequencies in images.

Numerous measures of histogram similarities exist (e.g. [22]) but not all of them are applicable to BoW matching. Because of the assumptions applied in this work during BoW building, we eventually selected a simple *histogram intersection* measure (proposed in [23]), where the distance between two histograms H_A and H_B over *Voc* vocabulary is defined by

$$d(H_A, H_B) = \sum_{w \in Voc} min(H_A(w), H_B(w)). \qquad (2)$$

Such a measure nicely corresponds to the intuitive notion of similarity between both full images and sub-images (including textured images).

B. Preliminary experimental results

The proposed approach was verified on UNICT-FD889 dataset discussed earlier. Not all dishes were fully tested, but we focused primarily on two most typical cases of food images. First, plates filled by uniformly looking dishes were considered (see examples in Fig. 2). Secondly, dishes represented by images with a number visually different regions of various food components (see Fig. 3) were taken into account.

The retrieval performances for both categories of dishes have been found very different. For uniform dishes, the top retrievals are usually highly relevant images. Images most similar to query images from Fig. 2 are shown in Figs 4 and 5, correspondingly. They highly correspond to the human perception, even though in Fig. 4 two categories of foods are mixed up.

For dishes consisting of several non-uniformly distributed items, performances are rather miserable. A spectacularly incorrect example is given in Fig. 6.

However, there are also cases (an example given in Fig. 7) when search results are quite good for multiple-item dishes.

IV. Concluding remarks

The paper proposes a novel CBVIR-based scheme for visual identification of food. Using the (apparently) largest publicly available dataset, we have tested a method based on BoW histograms which actually represent semi-local co-occurrences of features (MSER keypoints are selected as examples) extracted from three color channels of RGB images. Similar image retrieval is based on the similarities between such histograms.

Because of it low complexity, the scheme could be considered an attractive option for limited-performance mobile devices equipped with a camera.

Unfortunately, the experimental verification has been found only partially successful. The results are satisfactorily accurate

(a)

(b)

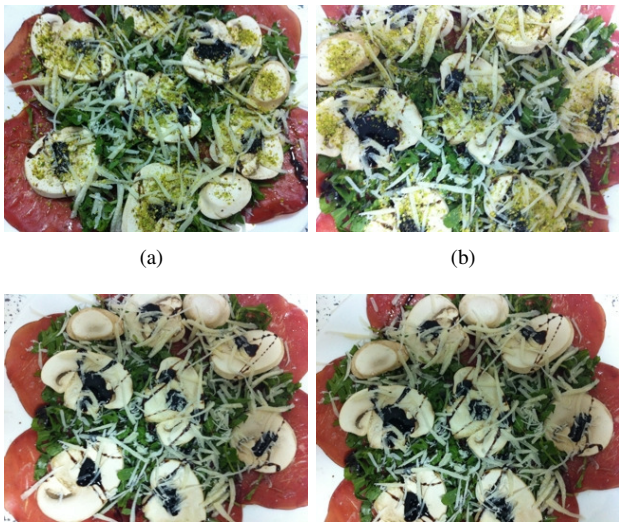
Fig. 2. Examples of dishes uniformly filling plates with similarly looking contents (from UNICT-FD889 dataset).

(a)

(b)

Fig. 3. Examples of dishes consisting of diversified components (from UNICT-FD889 dataset).

only for dishes looking uniformly over the whole plate. For mixtures of diversified foods shown on the same plate,



(a) (b)

(c) (d)

Fig. 4. Top retrievals for the image from Fig. 2a. Note that (a,b) and (c,d) are actually considered different dishes.

Fig. 5. Top retrievals for the image from Fig. 2b.

Fig. 6. Top retrievals for the image from Fig. 3a.

Fig. 7. Top retrievals for the image from Fig. 3b.

performances are unacceptably low (although in some cases perfomances are acceptable). Therefore, it can be preliminarily concluded that the approaches presented in some of the previous works, which require image segmentation into uniform regions (e.g. [3]) have been validated (in terms of the proposed methodologies).

However, those segmentation techniques proposed in the published works only partially correspond to needs identified in our experiments. For example, dishes shown in Fig. 2 should be considered uniform regions, but the existing segmentation technique (even if incorporating texture-based approaches) would apparently not segment them in the required way.

Altogether, wec can conclude that fully automatic vision-based food identification still remains a challenging problem. The main challenge is apparently segmentation of multiple-item dishes into uniform region, which can be prospectively recognized using keypoint-based approaches. Nevertheless the satisfactory solutions for such segmentation have not been identified yet. Keypoint-based co-segmentation (e.g. [24]) is one of the most promising approaches.

REFERENCES

[1] G. M. Farinella, D. Allegra, and F. Stanco, "A benchmark dataset to study the representation of food images," in *Proc. ECCV 2014 Workshops*, vol. III, 2015, pp. 584–599. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16199-0_41

[2] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.pmcj.2011.07.003

[3] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int.Conf. on Multimedia and Expo*, 2012, pp. 25–30. [Online]. Available: http://dx.doi.org/10.1109/ICME.2012.157

[4] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *Proc. IEEE Int. Symposium on Multimedia*, 2010, pp. 296–301. [Online]. Available: http://dx.doi.org/10.1109/ISM.2010.51

[5] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. CVPR 2010*, 2010, pp. 2249–2256. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2010.5539907

[6] Z. Zong, D. Nguyen, P. Ogunbona, and W. Li, "On the combination of local texture and global structure for food classification," in *Proc. IEEE Int. Symposium on Multimedia*, 2010, pp. 204–211. [Online]. Available: http://dx.doi.org/10.1109/ISM.2010.37

[7] G. O'Loughlin, S. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, and G. Warrington, "Using a wearable camera to increase the accuracy of dietary analysis," *American Journal of Preventive Medicine*, vol. 44, no. 3, pp. 297–301, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.amepre.2012.11.007

[8] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, 2010. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2010.2051471

[9] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," in *Proc. IEEE Conf. ICIP 2009*, 2009, pp. 289–292. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2009.5413511

[10] A. Jimenez, A. Jain, R. Ruz, and J. Rovira, "Automatic fruit recognition: a survey and new results using range/attenuation images," *Pattern Recognition*, vol. 32, no. 10, pp. 1719–1739, 1999. [Online]. Available: http://dx.doi.org/10.1016/S0031-3203(98)00170-8

[11] F. Pla, "Recognition of partial circular shapes from segmented contours," *Comput. Vision & Image Understanding*, vol. 63, no. 2, pp. 334–343, 1996. [Online]. Available: http://dx.doi.org/10.1006/cviu.1996.0023

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Conf. ICCV 2003*, vol. 2, Nice, 2003, pp. 1470–1477. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2003.1238663

[14] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996. [Online]. Available: http://dx.doi.org/10.1016/0031-3203(95)00067-4

[15] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-4635-4

[16] X. Qi, R. Xiao, J. Guo, and L. Zhang, "Pairwise rotation invariant co-occurrence local binary pattern," in *Proc. ECCV 2012*, 2012, pp. 158–171. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33783-3_12

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, pp. 761–767, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2004.02.006

[18] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans PAMI*, vol. 19, no. 5, pp. 530–535, 1997. [Online]. Available: http://dx.doi.org/10.1109/34.589215

[19] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. CVPR 2009*, 2009, pp. 25–32. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206566

[20] A. Śluzek, "Extended keypoint description and the corresponding improvements in image retrieval," *LNCS (Revised Selected Papers of ACCV 2014 Workshops)*, vol. 9008, pp. 698–707, 2015. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16628-5_50

[21] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. CVPR 2012*, 2012, pp. 2911–2918. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6248018

[22] S.-H. Cha and S. Srihari, "On measuring the distance between histograms," *Pattern Recognition*, vol. 35, pp. 1355–1370, 2002. [Online]. Available: http://dx.doi.org/10.1016/S0031-3203(01)00118-2

[23] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991. [Online]. Available: http://dx.doi.org/10.1007/BF00130487

[24] A. Śluzek and M. Paradowski, "Reinforcement of keypoint matching by co-segmentation in object retrieval: Face recognition case study," *LNCS (Proc. ICONIP 2012)*, vol. 7667, pp. 34–41, 2012. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34500-5_5

6th International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2016—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

- 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 Rzeszów, Poland

PROGRAM COMMITTEE

- **Azar, Ahmad Taher,** Benha University, Egypt
- **Iantovics, Barna,** Petru Maior University, Romania
- **Kountchev, Roumen,** Technical Univerity of Sofia, Bulgaria
- **Leniowska, Lucyna,** University of Rzeszow, Poland
- **Majernik, Jaroslav,** Pavol Jozef Safarik University in Kosice, Slovakia
- **Ngan, Ben C.K.,** The Pennsylvania State University, United States
- **Olszewska, Joanna Isabelle,** University of Gloucestershire, United Kingdom
- **Sawada, Hideyuki,** Kagawa University, Japan
- **Schwarz, Daniel,** Masaryk University, IBA, Czech Republic
- **Subbotin, Sergey,** Zaporizhzhya National Technical University, Ukraine
- **Wojciechowski, Konrad,** Silesian University of Technology, Poland
- **Zaitseva, Elena,** University of Zilina, Slovakia

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 65–70
DOI: 10.15439/2016F569
ACSIS, Vol. 9. ISSN 2300-5963

Automated 3D immunofluorescence analysis of Dorsal Root Ganglia for the investigation of neural circuit alterations: a preliminary study.

Santa Di Cataldo, Simone Tonti,
Enrico Macii, Elisa Ficarra
Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino (Italy)
Email: santa.dicataldo, simone.tonti@polito.it
enrico.macii, elisa.ficarra@polito.it

Elisa Ciglieri, Francesco Ferrini,
Chiara Salio
Dipartimento di Scienze Veterinarie
Universitá di Torino
Largo Paolo Braccini 2, 10095 Grugliasco (Italy)
Email: elisa.ciglieri, francesco.ferrini@unito.it
chiara.salio@unito.it

***Abstract*—Diabetic polyneuropathy is a major complication of diabetes mellitus, causing severe alterations of the neural circuits between spinal nerves and spinal cord. The analysis of 3D confocal images of dorsal root ganglia in diabetic mice, where different fluorescent markers are used to identify different types of nociceptors, can help understanding the unknown mechanisms of this pathology. Nevertheless, due to the inherent challenges of 3D confocal imaging, a thorough and comprehensive visual investigation is very difficult. In this work we introduce a tool, *3D*RG, that provides a fully-automated segmentation and 3D rendering of positively labeled nociceptors in a dorsal root ganglion, as well a quantitative characterisation of its immunopositivity to each fluorescent marker. Our preliminary experiments on 3D confocal images of entire dorsal root ganglia from healthy and diabetic mice provided very interesting insights about the effects of the pathology on two different types of nociceptors.**

I. Introduction

DIABETIC polyneuropathy (DPN) is one of the most common and serious complications of diabetes mellitus, which includes several types of nerve damaging disorders [1]. High glycemic levels associated with diabetes create injuries to the small vessels supplying the nerves, with symptoms that can range from pain and numbness in the extremities to problems with the digestive system, urinary tract, blood vessels and heart. Such symptoms in minor cases can be extremely disabling and even fatal.

While literature had traditionally focused only on injures of the peripheral nerves (mainly legs and feet), early works have now unveiled possible implications of DPN at all levels of the nervous system, with special regards to the neural circuits between the spinal nerves and the spinal cord [2]. Most recent studies are especially focusing on the Dorsal Root Ganglia (DRGs), clusters of sensory neurons in the dorsal root of spinal nerves (see Figure 1), whose underlying mechanisms in relation to DPN are at the moment poorly understood. In particular, the role of nociceptive sensory cells in the DRGs (i.e. neurons specialised in conveying pain information to the higher centers) is now one of the main topics of investigation [3], [4]. The analysis of immunofluorescence images via 3D confocal microscopy has a major role in such investigations. In particular, entire DRGs of mice can be dissected out and stained with multiple fluorescent markers, each targeting a specific type of nociceptor. The result is a complex multi-coloured stack of images, where different nociceptors are labelled by fluorochromes emitting signal of a known spectral range, hence they can be imaged in separate color channels (see left part of Figure 2). Typically used markers include biotin-conjugated Isolectin B4 (IB4) and the antibody for Calcitonin Gene-Related Peptide (CGRP), which identify the unmyelinated non-peptidergic and the small peptidergic neurons in the DRGs, respectively [3].

While the imaging technology per se is widely acknowledged for being a valuable support to this type of study, the analysis of 3D images of DRGs remains a challenging task. First, because distinguishing the positively stained neural cells is made difficult by the presence of noise and artefacts (e.g. spurious fluorescence, black spots, etc.), which are intrinsic limitations of immunofluorescence. Second, because the 3D nature of the images makes manual analysis unfeasible. To the best of our knowledge, there is no availability of a completely automated tool able to support this type of analysis. Hence, the data presented by most of the published works in this context are obtained with semi-automated procedures,

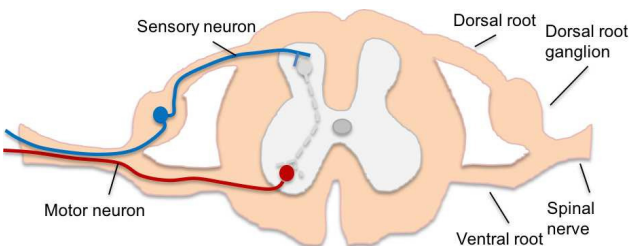


Figure 1. Cross-section of spinal cord.

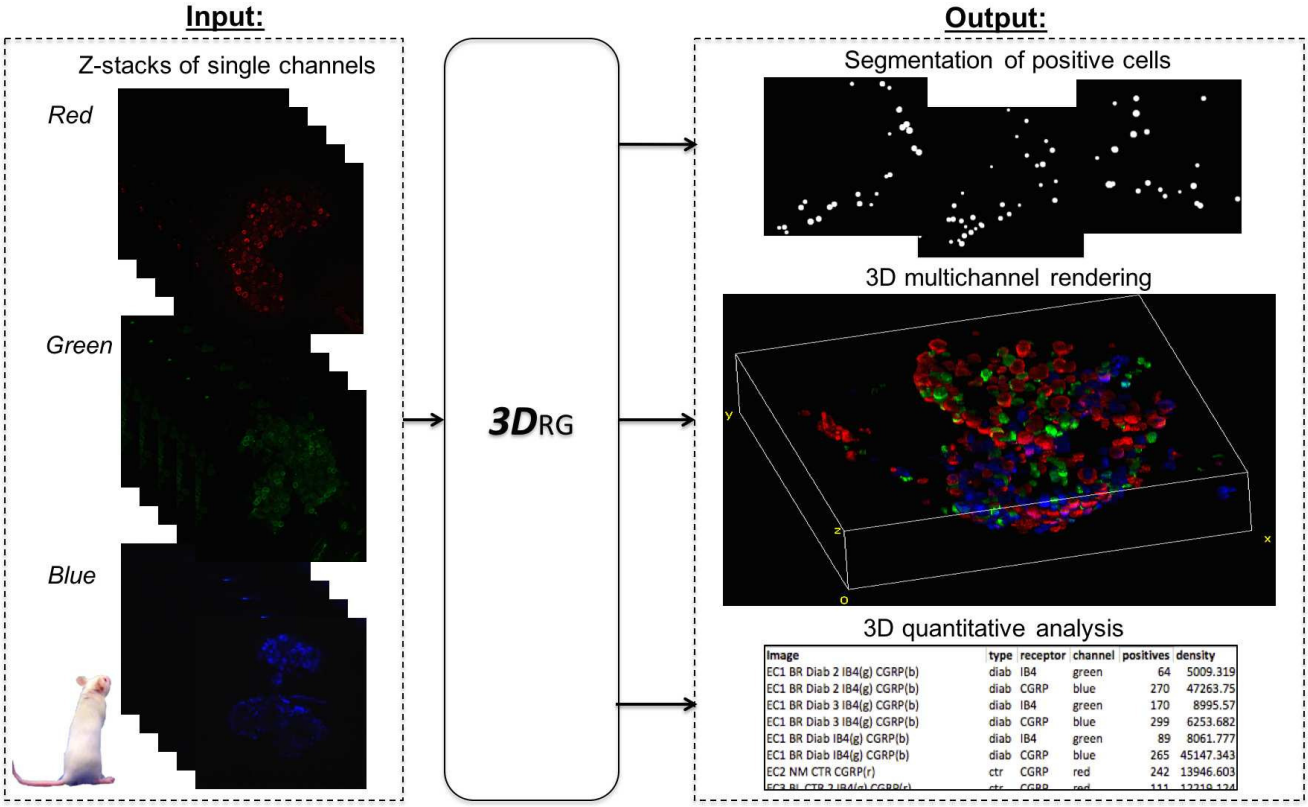


Figure 2. Simplified input/output diagram.

where the positive cells in few significant slices of the stack are distinguished from the background based on user-defined thresholds or standard 2D thresholding techniques. Then, the immunopositivity of the sample is quantified based on the number and average intensity of such cells [5], [3]. This process has two evident limitations: (i) it lacks reproducibility, due to the inherent subjectivity of user-interaction; (ii) by counting positive cells only on selected slices, it does take the 3D nature of the images into proper account. For example, a single neuron may span more consecutive slices of the stack, hence it should be counted only once.

As a solution to these problems, we propose *3D*RG, a tool for the immunofluorescence analysis of DRG samples. The main contribution of this work is two-fold. First, a fully automated 3D segmentation and 3D rendering of the positively labeled DRG neurons, based on a 3D spatial filtering technique. Second, a more accurate analysis of immunopositivity, which quantifies the number of positive cells and amount of fluorescent signal in the whole DRG stack.

Besides improving the feasibility and objectivity of DRG image analysis, *3D*RG allows the analysis of the spatial relations between cells marked by different types of antibodies. This opens up new perspectives in the investigation of the role and interaction of different types of nociceptor in the context of DPN alterations.

This paper is structured as follows. In Section II we provide the technical details of the three modules of our proposed tool. In Section III we characterise the DRG samples used in our preliminary experiments. In Section IV we report and discuss the obtained results. Finally, in Section V we draw conclusions and provide future perspectives of our work.

II. Proposed Method

As reported in Figure 2, our proposed tool *3D*RG receives as input the digitalised z-confocal stacks of the immunolabelled DRG samples, as returned by the confocal microscope. Separate modules of the tool provide the following output:

1) Segmentation of positive cells.
2) 3D multichannel rendering.
3) 3D quantitative analysis.

In the following, we describe the three modules in detail.

A. Segmentation of positive cells

The immunolabelled DRG cells are automatically detected in the 3D volume, discarding noise, spurious objects and artefacts. This is obtained by a 3D cell segmentation technique whose main steps are reported in the flow-charts of Figure 3.

1) Preprocessing: First, some preprocessing is performed on the 2D slices of the stacks, in order to ease the segmentation process by generally improving the image quality. The contrast between cells and background is enhanced by a combination of Contrast Limited Adaptive Histogram Equalisation technique [6] and background subtraction, where a simple model

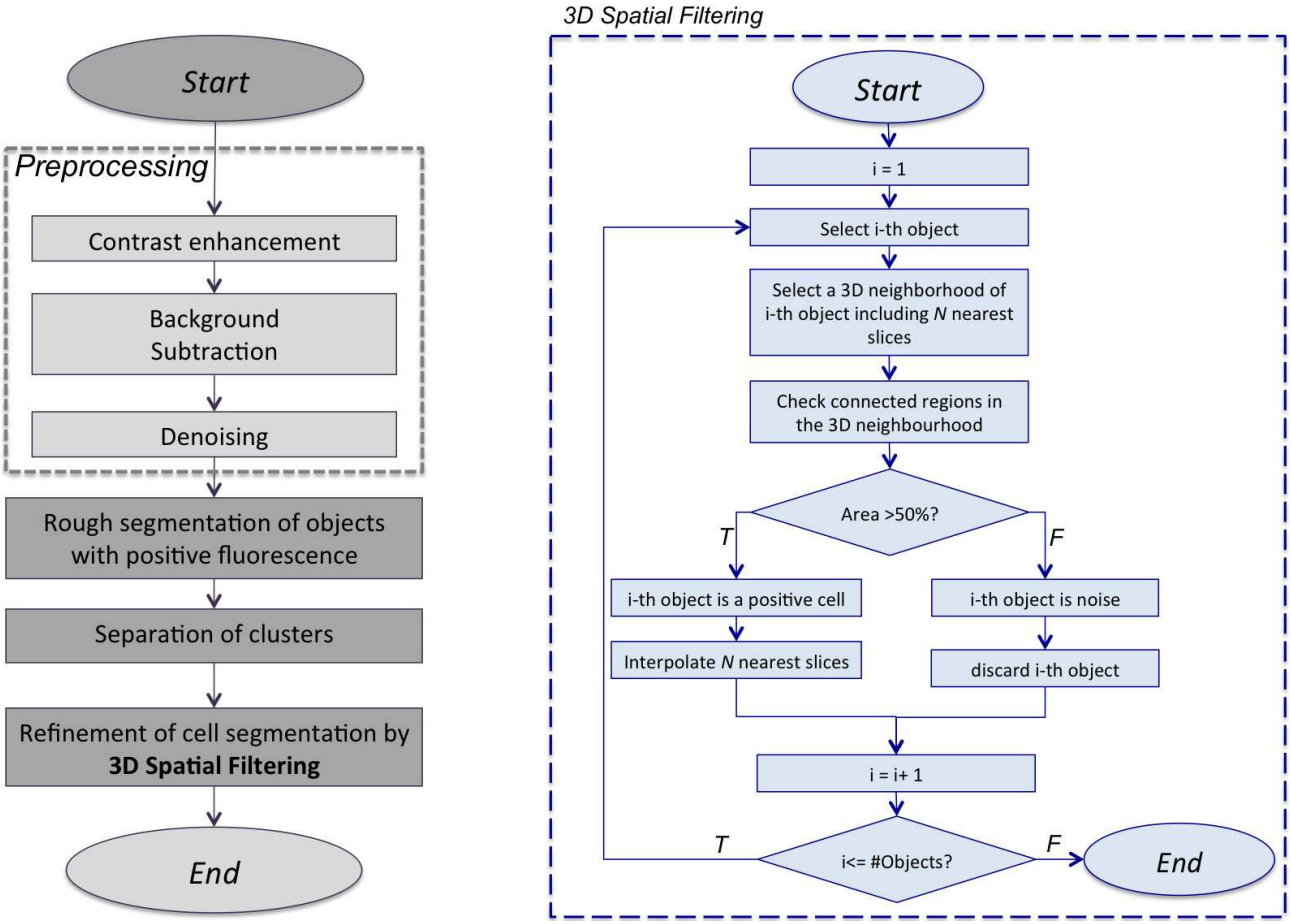


Figure 3. 3D Cell Segmentation: flow-chart.

of background computed by average filtering is subtracted to the original image. Following contrast enhancement, median filtering is performed to reduce high frequency noise preserving significant details of the image such as the borders of DRG cells.

2) Rough segmentation of objects with positive fluorescence: After preprocessing, fluorescent objects are roughly distinguished from the dark background by a 2D segmentation technique applied to each slice of the stack. This step implements a spatial fuzzy c-means algorithm (SFCM).

Standard fuzzy c-means (FCM) is a widely used clustering technique that partitions the data into a number of groups working towards the minimisation of the distance of points within the same cluster. The fuzziness of such process lies in assigning to the data points a $[0,1]$ membership level to each of the clusters (the so-called *membership functions*).

In order to decrease sensitivity to noise and reduce the spurious blobs, which are recurring problems in fluorescent image segmentation, *3D*RG uses a variant of conventional FCM that incorporates local spatial information into its implementation, by summing up the membership functions in the neighbourhood of each pixel [7]. After clustering, the shapes of the foreground objects are regularised by means of standard morphological operations such as opening and holes filling.

3) Separation of clusters: The foreground regions returned by SFCM algorithm may either contain one individual cell or multiple touching cells, which need to be separated into individual objects. Based on the assumption that individual cells are approximately circular, *3D*RG implements a Circle Hough transform (CHT), with the purpose of decomposing the input regions into a minimal number of circular components [8].

4) Refinement of cell segmentation: As ultimate step, the cell segmentation obtained by SFCM and CHT is refined by taking into account 3D spatial information (see **3D spatial filtering** flow-chart in Figure 3). More specifically, per each foreground object, the filter generates a 3D neighbourhood by projecting such object onto N consecutive slices of the z-stack (see Figure 4). N is set to be roughly equal to the expected length of DRG cells. Hence, its value depends on slice thickness.

If the 3D neighbourhood does not contain a significant portion (at least 50%) of positive regions, the corresponding object is interpreted as a spurious fluorescent spike, hence it is discarded. Otherwise, it is interpreted as part of a positive cell. In the latter case, the intensity values of consecutive slices in the 3D neighbourhood are interpolated along the z-axes, in

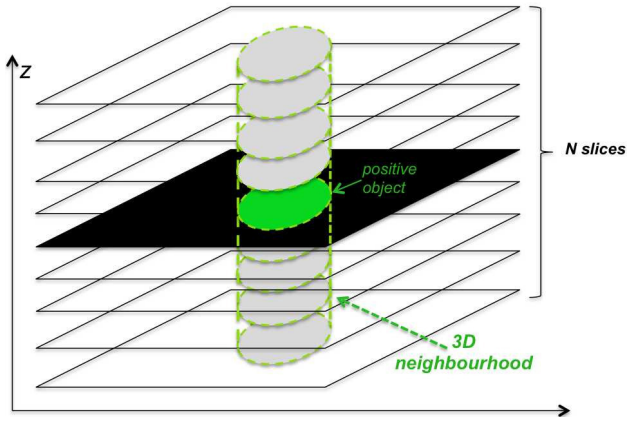


Figure 4. 3D neighbourhood for 3D spatial filtering.

order to remove black spots and ensure the 3D continuity of the cell.

More specifically, the pixels of the 3D neighbourhood in the slices above and below the positive object are replaced by the arithmetic mean of their corresponding pixels in the two closest slices, as follows:

$$slice(i) = \frac{slice(i-1) + slice(i+1)}{2}, \quad (1)$$

where i is the z-coordinate of the slice.

The final output of this step is a 3D volume where only the positive DRG cells have non-zero intensity values, proportional to the level of expression of the fluorescent marker.

B. 3D multichannel rendering

*3D*RG generates a volume rendering of the DRG-positive stack as returned by the cell segmentation module, with the possibility of displaying one channel at a time or more channels together (as in the example of Figure 2).

The user can interactively rotate the volume and visualise ortho-slices (i.e. three orthogonal slices through the volume). This allows to analyse spatial relations between different markers (e.g. cell clusters, recurring patterns, etc.) that are otherwise impossible to appreciate.

C. 3D quantitative analysis

The 3D sample is automatically characterized in terms of positive expression of each immunofluorescent marker.

This module allows the calculation of the following parameters:

- **Number of positive cells** (P), as automatically identified by the 3D cell segmentation algorithm.
- **3D Density** calculated as P/V_s, where P is the number of positive cells and V_S is the fraction of the sample volume with a non-zero fluorescent signal strength.
- **Integrated Optical Density (IOD)** calculated as the sum of mean intensity values (MIV) of the foreground (i.e. the positive cells) in each slice of the volume. More specifically:

$$IOD = \sum_{i=1}^{\#Slices} MIV_i, \quad (2)$$

where MIV_i is computed as the sum of the intensities of the foreground pixels in $i-th$ slice, divided by the total number of foreground pixels.
- **Related Optical Density (ROD)** calculated as:

$$ROD = IOD_F - IOD_B, \quad (3)$$

where IOD_F and IOD_B are the IOD values of the foreground and of the background, respectively.

III. Materials

In this work we performed experiments on DRG samples belonging to two populations of mice, respectively diabetic and healthy subjects (here referred as controls). Such samples were obtained as follows.

Four weeks old CD-1 male mice were made diabetic after a single intraperitoneal injection of streptozotocin (150 mg/Kg), while controls received only the vehicle. Glucose levels of all the subjects were weekly monitored, in order to ensure the correct categorization of diabetics and controls. All animals were sacrificed at eight weeks of postnatal age. Then, DRGs were acutely excised and the connective tissue was dissolved by incubation in 5-10 mg/mL collagenase. The entire DRGs were then used for immunofluorescence.

DRGs were stained for two classical phenotypic markers of nociceptors, i.e. the calcitonin gene-related peptide (CGRP) and the isolectin B4 (IB4) by a rabbit antibody and a biotin-conjugate, respectively. Whole DRGs z-stacks were then collected using confocal microscopy.

IV. Preliminary results

We run *3D*RG on a total number of 90 DRG samples. Of the 90 samples, 45 belonged to diabetic subjects and 45 to controls. Hence, the two populations were perfectly balanced.

Positive DRG cells were automatically detected as explained in Section II-A, and volume rendering of the segmented cells was generated (see the example of Figure 5).

IB4 and CGRP markers were automatically quantified in all the 90 samples, as reported in Section II-C.

The obtained results are summarised by the box-plots of Figure 6, grouping the data into diabetic and control subjects, respectively.

On each box-plot, the central red line is the median value, and the box has edges corresponding to the 25th and 75th percentiles (the so-called inter-quartile range IQR). The whiskers extend to the most extreme data points not considering outliers, while outliers are plotted individually using a red cross-shaped marker. As in previous works, all the values are normalised by the median of controls [3].

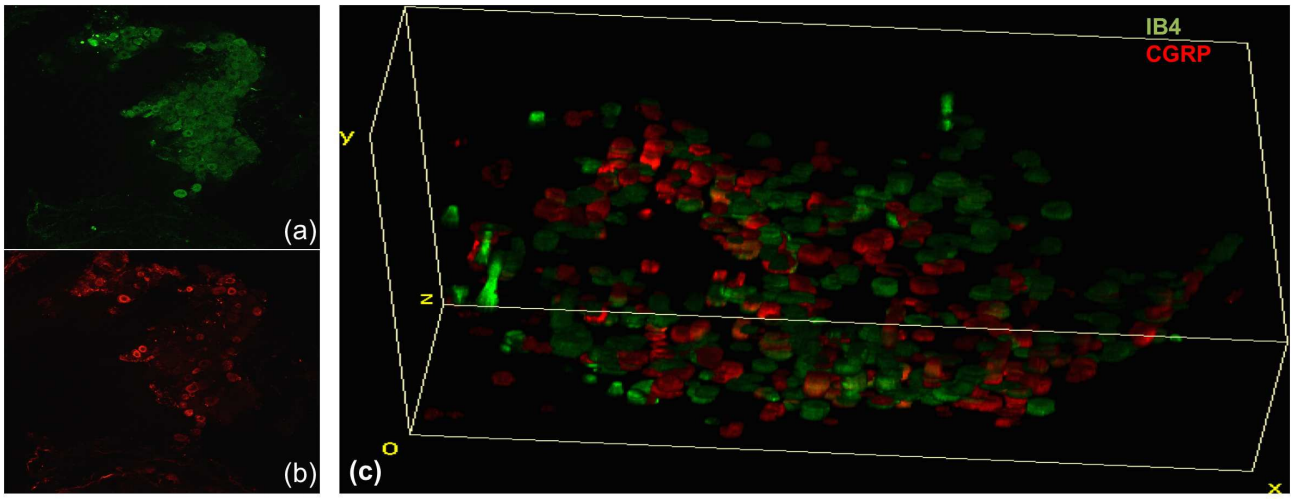


Figure 5. DRG sample. (a) IB4 labelling (green channel). (b) CGRP labelling (red channel). (c) Volume rendering of DRG positive cells automatically detected by *3D*RG tool.

In order to facilitate statistical comparisons between different groups, each box is displayed with a notch defining the confidence interval *C.I.* around the median, computed as:

$$CI = \text{median value} \pm 1.57 \cdot \frac{IQR}{\sqrt{N}}, \tag{4}$$

where N is the number of observations and the constant 1.57 is an empirical value that is set to approximate a 95% confidence interval around the median [9].

Hence, when the notches of two groups of data points do not overlap, it interpreted as a strong evidence that the medians of the two samples are significantly different.

From the analysis of the box-plots in Figure 6, the following considerations can be drawn:

1) 3D quantitative analysis revealed relevant differences between control and diabetic groups.
   In particular, the plots show a decrease of both IB4 and CGRP in the diabetic subjects.
2) the highest discrimination between controls and diabetics is obtained with IB4 marker. All four plots related to IB4 show non-overlapping notches between control and diabetic boxes, suggesting that the median values of the corresponding populations are different with a 95% confidence level.
   The same happens with CGRP, but only when considering IOD and ROD values.

The experimental results automatically obtained with *3D*RG are in line with the assumptions made by literature on neuroscience.

As reported by [3], [10], nonpeptidergic unmyelinated IB4-labeled afferents may have a higher susceptibility to diabetes, and their decrease might be a reason for the early sensory dysfunctions associated with this pathology.

On the other hand, CGRP-labeled peptidergic fibers are also to a lesser extent involved in the deficit.

V. Conclusions and Future Perspectives

In this paper we presented an automated tool, *3D*RG, that is able to

1) perform a segmentation of positively labeled DRG neurons;
2) provide a multichannel 3D rendering of the labeled neurons;
3) characterise the immunopositivity of the sample to the DRG markers.

Our proposed tool allows to obtain better insights into the analysis of immunofluorescence DRG images applied to the study of diabetic neuropathies, for two main reasons.

First, differently from previous works, where counting of positive cells was performed in a semi-automated way and on only few slices of the 3D stack, *3D*RG is able to characterise the sample in a fully-automated way, and by taking into account the whole 3D volume. This improves the repeatability and objectivity of the results, and allows to fasten and ease the analysis of large amount of image data.

Second, the 3D reconstruction and rendering of the segmented cells allows to visualise the 3D distribution of the different markers and to highlight spatial relations between different types of DRG afferents. This analysis cannot be performed on the original 3D stack due to noise and spurious fluorescence.

Results obtained in our preliminary experiments by running *3D*RG on DRG samples of healthy and diabetic mice were very interesting, in that they support the hypothesis that the alterations of the neural circuits between spinal nerve and spinal cord via the DRG might be involved in DPN, which is also confirmed by recent literature.

Indeed, fully-automated analysis of DRG images offers potential for huge improvements in the study of neural alterations related to diabetes. In our future work, we plan to extend *3D*RG to support a quantitative analysis of the 3D spatial

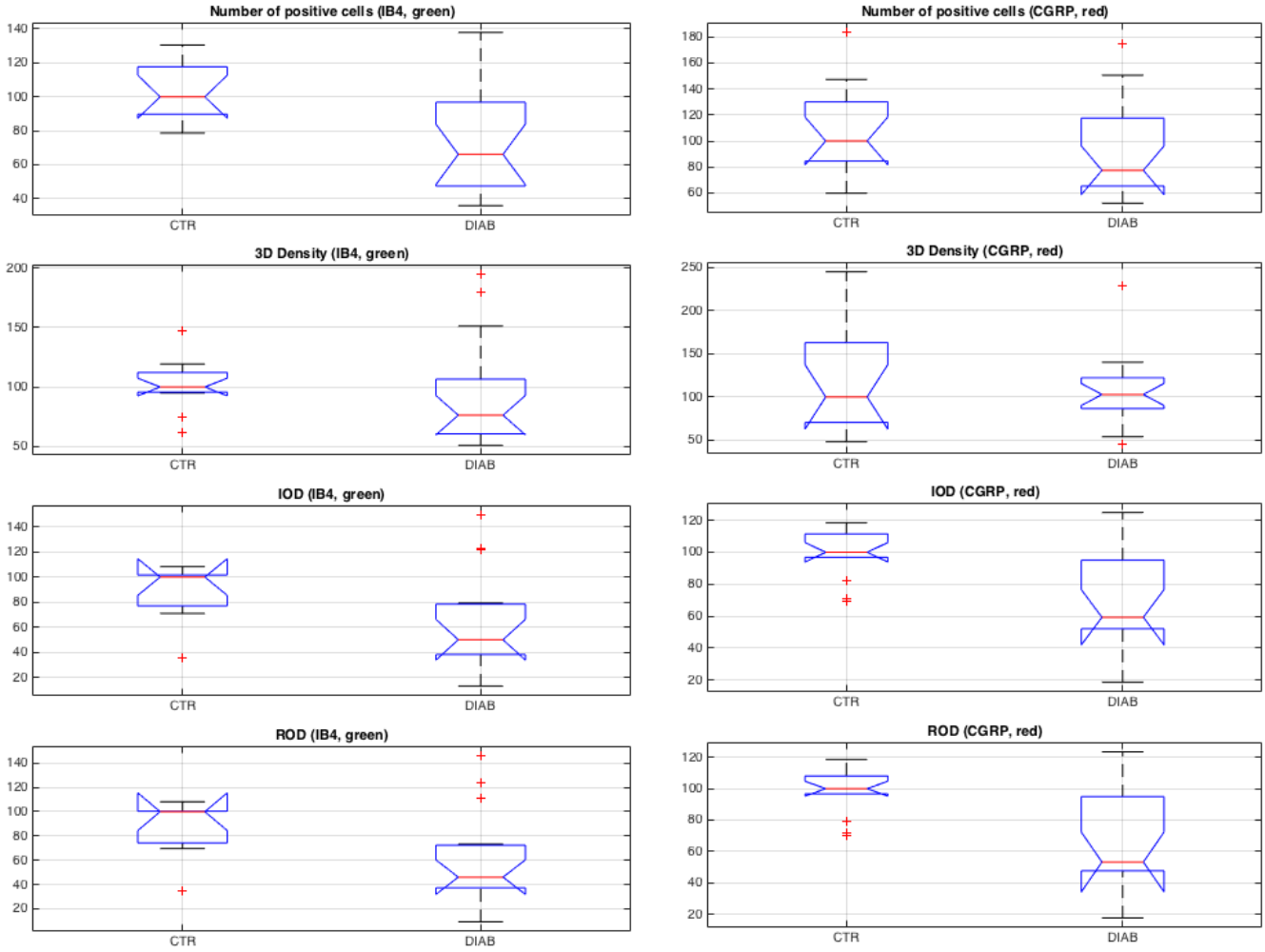


Figure 6. 3D quantitative immunofluorescence results on control and diabetic mice (values expressed as % of controls).

distribution of the different markers. This would allow to study the pathology-driven alterations of the relations between different DRG afferents, which is a type of analysis that was never performed before with immunofluorescence.

REFERENCES

[1] D. Ziegler and V. Fonseca, "From guideline to patient: a review of recent recommendations for pharmacotherapy of painful diabetic neuropathy," *Journal of Diabetes and its Complications*, vol. 29, no. 1, pp. 146 – 156, 2015. doi: 10.1016/j.jdiacomp.2014.08.008. [Online]. Available: http://dx.doi.org/10.1016/j.jdiacomp.2014.08.008

[2] H. Kamiya, K. E. Weixian Zhang and, J. Wahren, and A. A. Sima, "C-peptide reverses nociceptive neuropathy in type 1 diabetes," *Diabetes*, vol. 55, no. 12, pp. 3581 – 3587, 2006. doi: 10.2337/db06-0396. [Online]. Available: http://dx.doi.org/10.2337/db06-0396

[3] Z.-Z. Kou, C.-Y. Li, J.-C. Hu, J.-B. Yin, D.-L. Zhang, Z.-Y. Wu, T. Ding, J. Qu, Y.-H. Liao, H. Li, and Y.-Q. Li, "Alterations in the neural circuits from peripheral afferents to the spinal cord: possible implications for diabetic polyneuropathy in streptozotocin-induced type 1 diabetic rats," *Front Neur Circuits*, vol. 8, no. 6, 2014. doi: 10.3389/fncir.2014.00006. [Online]. Available: http://dx.doi.org/10.3389/fncir.2014.00006

[4] E. S. Krames, "The role of the dorsal root ganglion in the development of neuropathic pain," *Pain Medicine*, vol. 15, no. 10, pp. 1669–1685, 2014. doi: 10.1111/pme.12413. [Online]. Available: http://dx.doi.org/10.1111/pme.12413

[5] R. Wang, A. Rossomando, D. W. Sah, M. H. Ossipov, T. King, and F. Porreca, "Artemin induced functional recovery and reinnervation after partial nerve injury," *Pain*, vol. 155, no. 3, pp. 476 – 484, 2014. doi: 10.1016/j.pain.2013.11.007. [Online]. Available: http://dx.doi.org/10.1016/j.pain.2013.11.007

[6] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press Professional, Inc., 1994, pp. 474–485. [Online]. Available: http://dl.acm.org/citation.cfm?id=180895.180940

[7] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics*, vol. 30, no. 1, pp. 9 – 15, 2006. doi: 10.1016/j.compmedimag.2005.10.001. [Online]. Available: http://dx.doi.org/10.1016/j.compmedimag.2005.10.001

[8] T. Atherton and D. Kerbyson, "Size invariant circle detection," *Image and Vision Computing*, vol. 17, no. 11, pp. 795 – 803, 1999. doi: 10.1016/S0262-8856(98)00160-7. [Online]. Available: http://dx.doi.org/10.1016/S0262-8856(98)00160-7

[9] J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey, "Graphical Methods for Data Analysis," *The Wadsworth Statistics/Probability Series. Boston, MA: Duxury*, 1983. doi: 10.2307/2531418. [Online]. Available: http://dx.doi.org/10.2307/2531418

[10] S. Akkina, C. Patterson, and D. Wright, "{GDNF} rescues nonpeptidergic unmyelinated primary afferents in streptozotocin-treated diabetic mice," *Experimental Neurology*, vol. 167, no. 1, pp. 173 – 182, 2001. doi: 10.1006/exnr.2000.7547. [Online]. Available: http://dx.doi.org/10.1006/exnr.2000.7547

9th International 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.

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,** "Alexandru Ioan Cuza" University, Iasi, Romania
- **Chira, Camelia,** Technical University of Cluj-Napoca, Romania
- **Fidanova, Stefka,** Bulgarian Academy of Science
- **Gonçalves, Douglas,** Universidade Federal de Santa Catarina, Brazil
- **Hosobe, Hiroshi,** Hosei University, Japan
- **Iiduka, Hideaki,** Kyushu Institute of Technology, Japan
- **Lavor, Carlile,** IMECC-UNICAMP, Brazil
- **Marinov, Pencho,** Bulgarian Academy of Science, Bulgaria
- **Muscalagiu, Ionel,** Politehnica University Timisoara, Romania
- **Ninin, Jordan,** ENSTA-Bretagne, France
- **Parsopoulos, Konstantinos,** University of Ioannina, Greece
- **Pintea, Camelia,** Tehnical University Cluj-Napoca, Romania
- **Roeva, Olympia,** Institute of Biophysics and Biomedical Engineering, Bulgaria
- **Siarry, Patrick,** Universite Paris XII Val de Marne, France
- **Stefanov, Stefan,** South-West University ""Neofit Rilski, Bulgaria
- **Stuetzle, Thomas,** Université Libre de Bruxelles (ULB), Belgium
- **Tamir, Tami,** The Interdisciplinary Center (IDC), Israel
- **Zilinskas, Antanas,** Vilnius University, Lithuania

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 73–78
DOI: 10.15439/2016F168
ACSIS, Vol. 9. ISSN 2300-5963

Correlation clustering: divide and conquer

László ASZALÓS*, Mária Bakó†
* University of Debrecen
Faculty of Informatics
26 Kassai str., H4028 Debrecen, Hungary
Email: aszalos.laszlo@inf.unideb.hu
† University of Debrecen
Faculty of Economics
138 Böszörményi str., H4032 Debrecen, Hungary
Email: bakom@unideb.hu

***Abstract*—The correlation clustering is an NP-hard problem, hence its solving methods do not scale well. The contraction method and its improvement enable us to construct a divide and conquer algorithm, which could help us to clustering bigger sets. In this article we present the contraction method and compare the effectiveness of this new new and our old methods.**

I. Introduction

CLUSTERING is an important tool of unsupervised learning. Its task is to group objects in such a way, that the objects in one group (cluster) are similar, and the objects from different groups are dissimilar. It generates an equivalence relation: the objects being in the same cluster. The similarity of objects are mostly determined by their distances, and the clustering methods are based on distance.

Correlation clustering is an exception, it uses a tolerance (reflexive and symmetric) relation. Moreover it assigns to each partition (equivalence relation) a cost, i.e. number of pairs of similar objects that are in different clusters plus number of pairs of dissimilar objects that are in the same cluster. Our task to find the partition with the minimal cost. Zahn proposed this problem in 1965, but using a very different approach [1]. The main question is the following: *which equivalence relation is the closest to a given tolerance (reflexive and symmetric) relation?* Many years later Bansal at al. published a paper, proving several of its properties, and gave a fast, but not quite optimal algorithm to solve the problem [2]. Bansal have shown, that this is an NP-hard problem.

The number of equivalence relations of n objects, i.e. the number of partitions of a set containing n elements is given by Bell numbers B_n, where $B_1 = 1$, $B_n = \sum_{i=1}^{n-1} \binom{n-1}{k} B_k$. It can be easily checked that the Bell numbers grow exponentially. Therefore if $n > 15$, in a general case we cannot achieve the optimal partition by exhaustive search. Thus we need to use some optimization methods, which do not give optimal solutions, but help us achieve a near-optimal one.

If the correlation clustering is expressed as an optimization problem, the traditional optimization methods (hill-climbing, genetic algorithm, simulated annealing, etc.) could be used in order to solve it. We have implemented and compared the results in [3].

This kind of clustering has many applications: image segmentation [4], identification of biologically relevant groups of genes [5], examination of social coalitions [6], improvement of recommendation systems [7] reduction of energy consumption [8], modeling physical processes [9], (soft) classification [10], [11], etc.

In a previous paper [12] we presented the contraction method with many different interpretations, and later we constructed an improvement for the contraction method [13]. In this paper we introduce a new method which is based on the contraction method and its improvement, and this new method is a divide and conquer algorithm. By the measurements the new method is not much worse than the old one, therefore the new method could help us to solve concrete problems with thousands of objects.

The structure of the paper is the following:

In Section 2 we define correlation clustering mathematically and shortly present the contraction method and its improvement. Section 3 describes the divide and conquer algorithms. Next, we show the results of the measurements, and in Section 5 the recursive variant of the method. Later we present the results according to Barabási-Albert random graphs. In Section 7 we give our plans and discuss the technical details. Finally we conclude the results.

II. Correlation clustering

In the paper we use the following notations: V denotes the set of the objects, and $T \subset V \times V$ the tolerance relation defined on V. We handle a partition as a function $p : V \to \{1, \ldots, n\}$. The objects x and y are in a common cluster, if $p(x) = p(y)$. We say that objects x and y are in conflict at given tolerance relation and partition iff value of $c_T^p(x, y) = 1$ in (1).

$$c_T^p(x,y) \leftarrow \begin{cases} 1 & \text{if } (x,y) \in T \text{ and } p(x) \neq p(y) \\ 1 & \text{if } (x,y) \notin T \text{ and } p(x) = p(y) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We are ready to define the cost function of relation T according to partition p:

$$c_T(p) \leftarrow \frac{1}{2} \sum c_T^p(x,y) = \sum_{x<y} c_T^p(x,y) \tag{2}$$

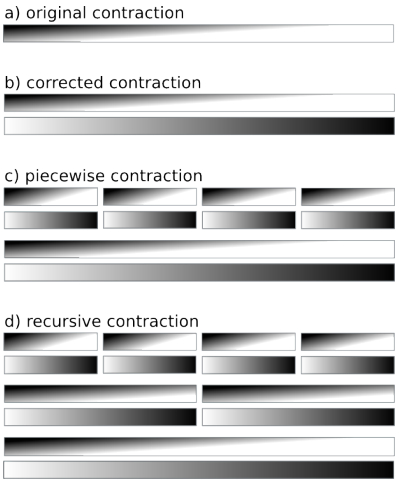


Fig. 1. Different variants of contraction method.

Our task is to determine the value of $\min_p c_T(p)$, and a partition p for which $c_T(p)$ is minimal. Unfortunately this exact value cannot be determined in practical cases, except for some very special tolerance relations. Hence we can only get approximative, near optimal solutions.

We can define the attraction between two objects: if they are similar then the attraction between them is 1; if they are dissimilar then the attraction between them is -1 (they repulse each other); otherwise—which can occur at a partial tolerance relation—the attraction is 0.

$$a(x,y) \leftarrow \begin{cases} 1, & \text{if } (x,y) \in T \\ -1, & \text{if } (x,y) \notin T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We can generalize (3) for object i and for clusters g and h:

$$a(x,g) = \sum_{y \in g} a(x,y) \text{ and } a(g,h) = \sum_{y \in h} a(y,g).$$

We leave it to the reader to check, that if these sums are positive and we join these element and clusters—by getting a partition p' containing the clusters $g \cup \{x\}$ or $g \cup h$—then $c_T(p) \geq c_T(p')$. This means that by joining attractive clusters, the cost decreases.

The contraction method starts with a partition where each cluster is a singleton. Next it selects clusters g and h for which $a(g,h)$ is maximal (and positive), and joins these clusters. It continues until there are no attractive clusters left. This contraction is presented on part a of Fig. 1 as a diagonally painted rectangle.

At dense Erdős-Rényi random graphs (ER in the following) each decision could have a vast impact, at the pure contraction method we cannot correct previous decisions—we just join, and not slit—, therefore we need something else in order to make a decision. This is the place, where the attraction between a node and a cluster is taken into consideration: we take every combination of nodes and clusters. If we find that a node is attracted much harder by any other cluster than by its clusters, then we put that node into the other cluster. Moreover if some node is repulsed by all the clusters, we construct an extra cluster for this node and we move it into this new cluster. This is the improvement of the contraction, that corrects the faults of the contraction. This correction step on part b of Fig. 1 denoted as a vertically painted rectangle. This correction step follows the contraction step. Here we apply both steps for the whole V i.e. for all objects together.

We discussed the properties of the contraction and this improvement in [13].

III. Divide and conquer algorithms

The divide-and-conquer strategy solves a problem by:

- breaking it into sub-problems that are themselves smaller instances of the same type of problem—*divide*
- recursively solving these sub-problems—*conquer*
- appropriately combining their answers—*combine*

One may think that only the second step (conquer) could be hard, in general each step needs some work: remember the splitting of the array at quick-sort for the divide step, and the merging of the ordered sequences for the combine step of the merge sort.

As practice has shown, we can amend the outcome of the contraction with its improvement, so we treat the contraction and its improvement as a unit. This is the reason why part c and d of Fig. 1 are built from units (of contraction and correction) of part b.

At a correlation clustering problem we have a set V of n objects, so the division is simple: split the whole set into two subsets of $n/2$ objects (or three subsets of $n/3$ objects, and so on).

The subsets are similar to the original sets, so at the conquer step we only need to apply the unit for each subset independently, as first half of part c of Fig. 1 presents.

The outcome of correlation clustering is a partition, so the outcome of the correlation clustering of the subsets are independent partitions as Fig. 2 shows. At the combine step of the algorithm we need to check whether the clusters of the partitions of the subsets could be parts of a cluster, which is member of the partition of the whole set. Fortunately we have a tool to answer this question, the contraction does exactly this: checks whether two cluster could be joined. Hence we need to apply the unit of contraction and correction for the whole set V, as the second half of part c of Fig. 1 presents.

Based on the algorithm of the contraction (`contract`) and its improvement (`correct`) we can easily implement in Python this divide and conquer algorithm, as Alg. 1 shows. Here `uf` is the UnionFind data structure that handles the data of the partition. `self.size` stores the number of objects in V, and `no_parts` denotes the number of subsets. The contraction and its improvement were implemented in such a way, that they get the subsets of the objects with their (`lower` and `upper`) bounds. Whilst they only use one data structure to store attraction between clusters and objects, they handle these data independently for different subsets of objects.

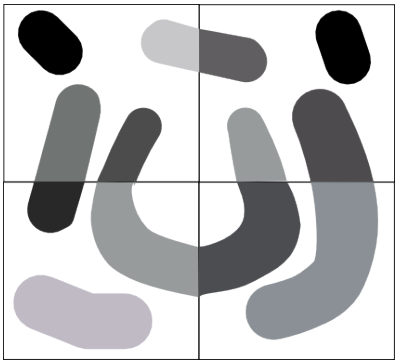


Fig. 2. Outcome of partitioning the subsets

Algorithm 1 Corrected contraction

```
def contract_in_pieces(self, uf, no_parts):                1
    sub_size = math.ceil(self.size/no_parts)               2
    for lower in range(0, self.size, sub_size):            3
        upper = min(lower + sub_size, self.size)           4
        self.contract(uf, lower, upper)                    5
        self.correct(uf, lower, upper)                     6
    self.contract(uf, 0, self.size)                        7
    self.correct(uf, 0, self.size)                         8
```

In the last two lines of Alg. 1, we combine the clusters of the subsets by applying the contraction to the whole set of objects. (We note that Python starts indexing with 0, and at defining intervals, the upper limit is exclusive.)

The implemented methods and the testing environment is available at https://github.com/aszalosl/DC-CC.

IV. Effectiveness of the new method

At the first measurements we use traditional tolerance relations, i.e. the signed graph of the problem is total. The structure of the relation—or the structure of the graph—determines the cost $c_T(p)$, e.g. despite that two graphs have the same number of negative and positive edges, one graph belongs to an equivalence relation—so its cost is 0—, while the cost of the other graph is a large number. It is extremely unlikely that we will obtain an equivalence relation by randomly generating a tolerance relation, and by our former experiments related to total graphs the deviation of the costs is small.

Although the number of positive and negative edges does not precisely describe the graph (and hence the relation), it helps us to understand the processes. We will use the rate calculated by the number of positive edges and the number of all edges, and denote this ratio by q. On Fig. 3 we present rates of different $c_T(p)$'s as a function of q (of T), where the p is the partition that the algorithm generates from T. We used 100 random relations with the same q, and we display the mean here. The base is the contraction method without any improvement/correction, this belongs to the level 1.0. The solid line denotes the contraction with its improvement. As the solid line does not exceed the level 1.0, it really is an improvement.

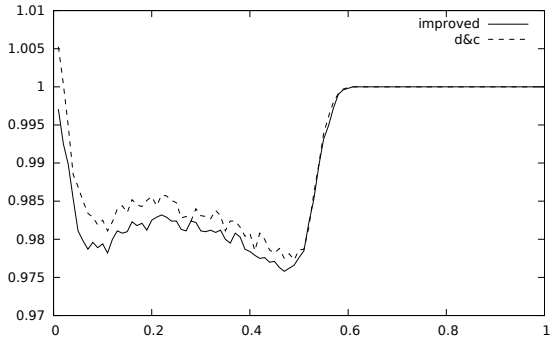


Fig. 3. Comparison of the original, the improved and the divide and conquer method with 10 subsets. (Less is better.)

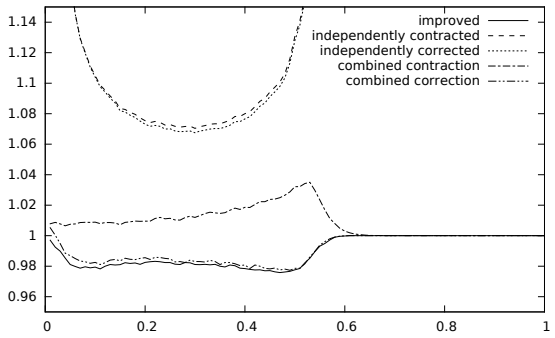


Fig. 4. Results of different steps of the piecewise contraction.

The dashed line denotes the outcome of the divide and conquer algorithm when we split the original problem into ten sub-problems. As it is slightly above the solid line, it almost reaches the level of the improved algorithm.

Fig. 4 goes into details. The solid line below shows the result of the corrected contraction, where we applied both process for the whole set V. On the other hand at the fist stage of the piecewise contraction we need to execute the contraction (in parallel) for the subsets of V. As the graph of the problem is a total graph, i.e. everything is connected, the attraction of the objects in different subsets suffers from cost function, so its dashed line is at the top. The piecewise correction (second stage) is an improvement, so its dotted line is below the previous line. The clusters of the partitions of the subsets could be joined at the next stage, which a total contraction—we apply it for the whole set V—and with this we reach the level 1.0 somewhere. And the last stage—the total correction—get below this level, and approaches the improved contraction.

A huge sample could smooth the lines, but we do not believe that using big samples gives clearly answer which parameter is the best for Fig. 5. Here we compare the results of dividing the original set into different number of subsets. The lines intersect each-other several times on the graph, therefore we cannot announce the absolute winner, maybe the biggest parameter gives the best (smallest) result.

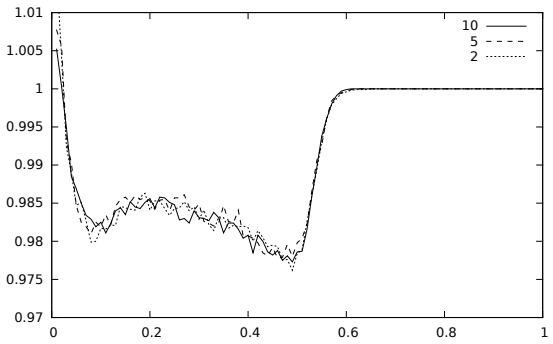


Fig. 5. Comparison of the divide and conquer method with different number of subsets.

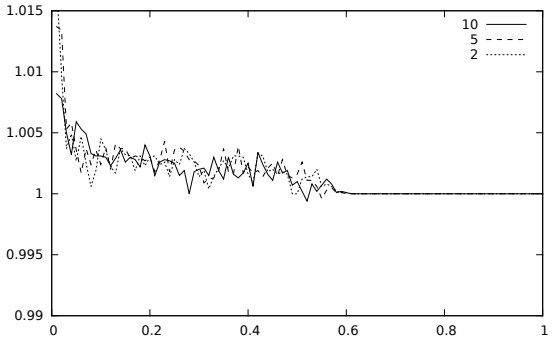


Fig. 6. Comparison of the divide and conquer method with different number of subsets.

Does the improved or the divide and conquer method give better result? We created Fig. 6, to show the rate of their result in each case. As the lines are mostly above the level 1.0, the divide and conquer method gives a bit worse result, than the improved algorithm. This is a reasonable result: if we have less nodes, we only know half of the information and hence our decisions are less sound and we make more mistake. We tested 100 complete graphs each with 200 nodes and divided them into 10-element sub-graphs, and when we compared the result, strangely the divide and compare was better with one percent than the improved contraction alone.

If we allow partial tolerance relations, the diversity increases. Fig. 7 uses the same number of objects, but the graph of the relation is an ER graph with different ps. As there are less edges left, they could create a more complex structure than in a total graph, so the variance is bigger than before, and this increases as the parameter p decreases. However, the situation itself is similar, so the improved method is slightly better than the divide and compare one.

V. Recursive algorithm

If we want to cluster 10^6 objects, it does not help us to cluster ten set of objects with 10^5 members in each, although previous experiments suggest that less subsets are better. Most of the divide and conquer algorithm uses recursion, let us apply it for this problem! Alg. 2 receives an UnionFind structure

Algorithm 2 Recursive contraction

```
def rec_c(self, uf, l, lower, upper):                1
    if lower + l < upper:                            2
        mid = (lower + upper)//2                     3
        self.rec_c(uf, l, lower, mid)                4
        self.rec_c(uf, l, mid, upper)                5
    self.contract(uf, lower, upper)                  6
    self.correct(uf, lower, upper)                   7
```

to store the clustering, the maximal length l of a primitive cluster—that will not be divided further—, and the bounds of the subset. If the size of the cluster is bigger then l, we divide it into half, and recursively call both parts. Next, we apply the contraction, and later the correction for the whole subset.

Last part of Fig. 1 shows, that we apply the contraction and corrections for quarter sets, next on the half sets, and finally the whole set. The parameter l enables us to try several values, as Fig. 8 shows. It may be hard to read from the picture, but we believe that smaller values give smaller cost, hence are better. We have calculated the sum of the rates for a big sample and this sum is less for smaller ls, so we get closer to the result of the improved contraction from above. Taking a look at the last part of Fig. 1, we can say, that is contains numerous corrections. Can we omit all of them but the last? (We need to delete Line 7 from Alg. 2, and put it into the main program.) Yes, we can, but the result is poor as Fig. 8 shows.

The difference between piecewise and recursive contraction is small, hence further research is needed to decide which one could help us to cluster large sets.

VI. Sparse random graphs

Until now we examined ER graphs. In nature, this kind of graph is not common, so we take a look at other types of graphs. The other well-known random graph type is the Barabási-Albert (BA in the following). At generating this type of graphs we use preference attachment. Here each new node is connected to existing nodes with a probability that is proportional to the number of links that the existing nodes already have. This means, that the oldest nodes construct a relatively dense graph, and the young nodes connect them

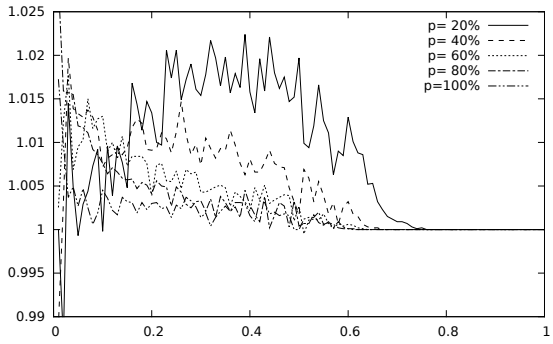


Fig. 7. Contacting Erdős-Rényi random graphs with different probabilities.

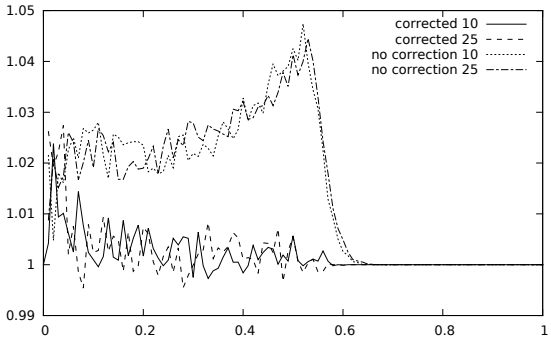


Fig. 8. Different lengths at recursive contraction and at a variant.

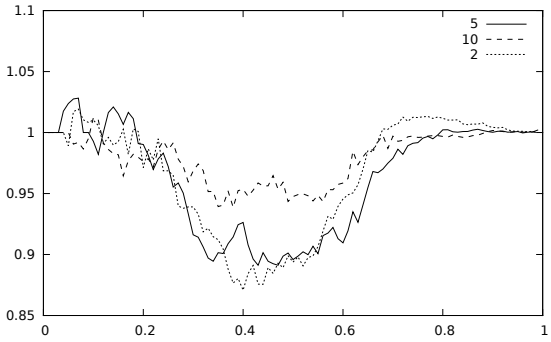


Fig. 9. Different number of subsets at clustering BA graphs.

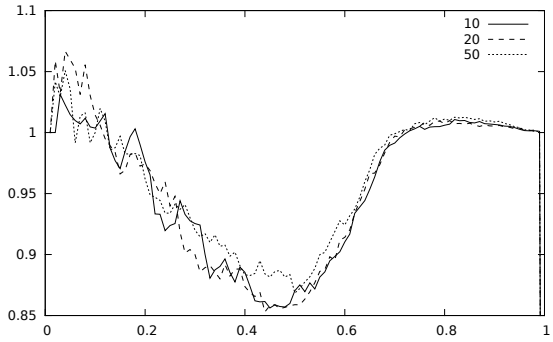


Fig. 10. Different primitive sizes at clustering BA graphs.

TABLE I
RUNNING TIME OF DIVIDE AND CONQUER CLUSTERINGS.

size	number of parts				max size of primitive cluster			
	2	5	10	20	10	20	25	50
ER graphs								
50	0.82	0.74	0.72	0.73	0.74	0.76	0.81	0.96
100	0.69	0.59	0.56	0.55	0.56	0.57	0.61	0.69
200	0.52	0.39	0.35	0.33	0.34	0.35	0.37	0.41
BA graphs								
500	0.92	0.91	1.08	1.20	0.78	0.77	0.77	0.76
1000	0.91	0.90	1.07	1.21	0.74	0.73	0.73	0.73
2000	0.89	0.89	1.06	1.21	0.71	0.70	0.70	0.70

weakly. By slicing the set of objects into subsets based on their (serial) numbers, we put the oldest nodes into the same subset. Applying our methods to these kind of subsets we get very similar results as before. To discover the real tendencies, we mixed up the numbers of the objects, and used the simplest slicing.

Our previous research had shown that the clustering of ER type and BA type random graphs gives different results and different tendencies [13]. Therefore we repeated our previous tests for these graphs, too. As sparse graphs have only a few edges, here we used 500 objects and 3/2 type BA graphs. Fig. 9 shows that strange finding, that the cost for five subset is minimal, meaning that for two or ten subsets the cost is higher. We have used different samples in this tests, hence we need test the same graphs to get more reliable results. Fig. 10 shows that smaller primitive subsets give better result, but we tested this on different samples too. The variance is big at BA graphs, so it is better to use the same graphs here.

Some tendencies are the same as at ER types, but the rate of the improved contraction and piecewise/recursive contraction is the opposite. At ER graphs the improved contraction is slightly better, but at BA graphs the the latter methods produce significantly better results.

VII. TECHNICAL DETAILS

We have not yet mentioned the speed of the methods. Of course, the parameter q influences the speed. If q is low, then there is little chance of contraction, so at the combine step of the algorithm we have almost the same number of clusters as at the beginning. This means, that we do the same task twice. If q is high, then most of the objects are contracted into few clusters, and contracting them does not take much time. At the clustering of the subsets we need to cope with smaller complexity, therefore we hope that at high q-s the divide and conquer algorithm could run faster than originally.

We did not perform very detailed analysis for each value of q, but we calculated the mean of the running time with different parameters. Even this shows the tendencies, and determines the future research directions.

Table I shows the rates of running times of the different divide and conquer algorithms. The base is the running time of the improved contraction method, and if the number is less than 1.0, then the divide and conquer algorithm is faster.

At the two different kinds of random graphs the algorithms behave differently. At ER graphs as we divide the original set into more and more parts, the running time decreases. At BA graphs more and more parts require extra overhead, and eventually its running time becomes higher than the original method's.

If we use a recursive method, the calculations speed up at ER graphs with smaller sets. At two hundred nodes, we can reduce the running time to its third. We hope, that at bigger sets we can save even more time. At BA graphs we can reduce

the running time too. Unfortunately, it does not speed up as much as the ER graphs. Here the more level of recursion does not help, moreover it has some overhead, so the running time is longer than at smaller level of recursions.

Do not forget, that BA graphs have $O(n)$ edges, while ER graphs have $O(n^2)$ edges, and this property holds for its subgraphs. To work with $k \cdot c \cdot \left(\frac{n}{k}\right)^2$ instead of $c \cdot n^2$ is better, while there is no real difference $k \cdot c \cdot \left(\frac{n}{k}\right)$ and $c \cdot n$.

The former property holds for ER graphs with probability $p \neq 1.0$, hence the rate running time of divide and conquer clusterings is very close to the numbers in the table.

For UnionFind data structures there is an excellent implementation: disjoint-set forests. Unfortunately this implementation does not contain the *replace* operator, which is necessary to correct the errors of the contractions. There are algorithms which enable the deletion [14], but this is a logical deletion and not a physical one, which would be needed to replace the old value with a new one.

Our implementation currently uses an array assigning the identifier of a cluster to each object, and an associative array assigning the cluster to an identifier. In this case, the operation union have linear complexity according to size of the smaller cluster, and the other operations have constant complexity.

Our implementation caches values $a(x, g)$ and $a(h, g)$ for every valid x, g and h. As these values change at contraction and at correction; we need to update the stored values. This requires many small tricks, but is much faster than the version calculating these values again and again.

VIII. Conclusion and further work

Based on our contraction method and its improvement we constructed a divide and conquer algorithm. Our hypothesis was that it gives poor results, because it uses less information, than the original method, but in some cases runs faster than the original algorithm. Surprisingly, the new method generates result close the improved variant of the former method at ER graphs, and gives better result for BA graphs. The former data structures and its methods are not usable for us, so we applied less sophisticated algorithms. The similar results and a faster calculation suggests, that this could be a fruitful direction. We need a more detailed comparison of methods, to discover the limits of this algorithm; and extra work to implement the parallel version, which uses the advantages of the divide and conquer method.

References

[1] C. Zahn, Jr, "Approximating symmetric relations by equivalence relations," *Journal of the Society for Industrial & Applied Mathematics*, vol. 12, no. 4, pp. 840–847, 1964. [Online]. Available: http://dx.doi.org/10.1137/0112071

[2] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:MACH.0000033116.57574.95

[3] L. Aszalós and M. Bakó, "Advanced search methods (in Hungarian)," http://morse.inf.unideb.hu/~aszalos/diak/fka, 2012.

[4] S. Kim, S. Nowozin, P. Kohli, and C. D. Yoo, "Higher-order correlation clustering for image segmentation," in *Advances in Neural Information Processing Systems*, 2011, pp. 1530–1538.

[5] A. Bhattacharya and R. K. De, "Divisive correlation clustering algorithm (dcca) for grouping of genes: detecting varying patterns in expression profiles," *bioinformatics*, vol. 24, no. 11, pp. 1359–1366, 2008. [Online]. Available: dx.doi.org/10.1093/bioinformatics/btn133

[6] B. Yang, W. K. Cheung, and J. Liu, "Community mining from signed social networks," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 10, pp. 1333–1348, 2007.

[7] T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan, "Improving recommendation accuracy by clustering social networks with trust," *Recommender Systems & the Social Web*, vol. 532, pp. 1–8, 2009. [Online]. Available: http://dx.doi.org/10.1145/2661829.2662085

[8] Z. Chen, S. Yang, L. Li, and Z. Xie, "A clustering approximation mechanism based on data spatial correlation in wireless sensor networks," in *Wireless Telecommunications Symposium (WTS), 2010*. IEEE, 2010, pp. 1–7. [Online]. Available: http://dx.doi.org/10.1109/WTS.2010.5479626

[9] Z. Néda, R. Florian, M. Ravasz, A. Libál, and G. Györgyi, "Phase transition in an optimal clusterization model," *Physica A: Statistical Mechanics and its Applications*, vol. 362, no. 2, pp. 357–368, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2005.08.008

[10] L. Aszalós and T. Mihálydeák, "Rough clustering generated by correlation clustering," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Springer Berlin Heidelberg, 2013, pp. 315–324. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.1061

[11] L. Aszalós and T. Mihálydeák, "Rough classification based on correlation clustering," in *Rough Sets and Knowledge Technology*. Springer, 2014, pp. 399–410. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11740-9_37

[12] L. Aszalós and T. Mihálydeák, "Correlation clustering by contraction," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015, pp. 425–434.

[13] L. Aszalós and T. Mihálydeák, "Correlation clustering by contraction, a more effective method," to appear in Studies in Computational Intelligence.

[14] S. Alstrup, M. Thorup, I. L. Gørtz, T. Rauhe, and U. Zwick, "Union-find with constant time deletions," *ACM Transactions on Algorithms (TALG)*, vol. 11, no. 1, p. 6, 2014.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 79–85
DOI: 10.15439/2016F414
ACSIS, Vol. 9. ISSN 2300-5963

Testing of parallel metaheuristics for graph partitioning problems

Zbigniew Kokosiński, Pawel Bala
Cracow University of Technology,
Faculty of Electrical and Computer Eng.
ul. Warszawska 24, 31-155 Kraków, Poland;
Email: zk@pk.edu.pl

***Abstract*—In this paper we describe computer experiments while testing a family of parallel and hybrid metaheuristics against a small set of graph partitioning problems like clustering, partitioning into cliques and coloring. In all cases the search space is composed of vertex partitions satisfying specific problem requirements. The solver application contains two sequential and nine parallel/hybrid algorithms developed on the basis of SA and TS metaheuristics. A number of tests are reported and conclusions resulting from the testing experiments are derived.**

***Index Terms*—simulated annealing, tabu search, parallel metaheuristic, hybrid metaheuristic, graph partitioning problem**

I. Introduction

COMPUTATIONAL optimization attracts for years researchers and practitioners interested in solving combinatorial problems by means of various computational methods and tools. In particular, many NPO problems require new versatile tools in order to find approximate solutions [1], [8]. Parallel and hybrid metaheuristics are among the most promisssing methods to be developed in the nearest time [2], [16]. Many new algorithms have been already designed and compared with existing methodologies [7], [11], but there is still a room for significant progress in this area.

In this paper we focus on a class of partitioning problems that appears in many application areas like data clustering [3], column-oriented database partitioning optimization [15], design of digital circuits, decomposition of large digital systems into a number of subsystems (moduls) for multi-chip implementation, task scheduling, timetabling, assiggnment of frequencies in telecommunication networks, etc. Partitionig problems are in general simpler than permutation problems but their search spaces are too huge for exhaustive search or extensive search methods [6], [9], [10], [17], [18].

The rest of the paper is organized as follows. In the next section the graph partitioning problems are defined and characterized. Then, in section 3, SA and TS algorithms as well as their parallelization and hybridization methods are presented. The design assumptions and features of the developed solver are described in section 4. Testing methodology and experimental results are shown in section 5. The final conclusions point out the directions of future research in this area.

II. Graph partitioning problems

In this section formulations of several partitioning problems are given that are to be solved by a collection of algorithms used in the experimental part of the paper.

We assume that $G = (V, E)$ is a connected, undirected graph. Let $|V| = n$, $|E| = m$.

A. Cluster partitioning problem (CPP)

A partition $C = (C_1, \ldots, C_k)$ of V is called a clustering of G and C_i clusters. C is called trivial if either $k = 1$, or all clusters C_i contain only one element. We will identify a cluster C_i with the induced subgraph of G, i.e. the graph $G_i = (C_i, E(C_i))$, where $E(C_i) = \{\{u, v\} \in E : u, v \in C_i\}$. Hence, $E(C) = \sum_{i=1}^{k} E(C_i)$ is the set of intra-cluster edges and $E \setminus E(C)$ the set of inter-cluster edges. [3]

The number intra-cluster edges is denoted by $m(C)$ and the number of inter-cluster edges by $M(C)$.

The $coverage(C)$ of a graph clustering C is a fraction of intra-cluster edges within the complete set of edges E: $coverage(C) = m(C)/m$. The larger the value of $coverage(C)$ does not necessarily mean the better quality of a clustering C.

Constructing a k-clustering with a fixed number of k, $k \geq 3$ of clusters is NP-hard [1].

In this paper we will consider k-clustering problems for weighted graphs, where the total weight of the set $E \setminus E(C)$ shall be minimized.

B. Clique partitioning problems (CPP)

A partition $C = (C_1, \ldots, C_k)$ of V is called a partition of G into cliques iff every subgraph $G_i = (C_i, E(C_i))$ induced by a cluster C_i is a clique, i.e. all vertices in C_i are pairwise connected. The goal is to find the minimal k, for which a partition into at most k cliques exists.

The clique partitioning problem is NP-complete [14]. The dual problem to CPP is graph partitioning into independent sets (ISs). It is equivalent to the CPP for $G(V, E')$, where E' is a complement of the set E.

C. Clique partitioning problems with minimum clique size (CPP)

In the present paper a solution of clique partitioning problem is also searched for given clique size at least s: is there a

graph partition into k cliques satisfying a condition related to the minimum clique size s? For given n and k the minimum size of cliques in G is $s = \lfloor n/k \rfloor$. Weighted version of the problem are also known, with additional conditions related to cliques' weights [9].

D. Graph coloring problem (GCP)

Classical vertex coloring problem in a graphs is another formulation of graph partitioning into independent sets. Such ISs can be assigned different colors, satisfying the property that all pairs of adjacent vertices in G are assigned nonconflicting colors. Formally:

For given graph $G(V, E)$, the optimization problem GCP is formulated as follows: find the minimum positive integer k, $k \leq n$, and a function $c : V \longrightarrow \{1, \ldots, k\}$, such that $c(u) \neq c(v)$ whenever $(u, v) \in E$. The obtained value of k is referred to as graph chromatic number $\chi(G)$.

GCP belongs to the class of NP-complete problems [8].

E. Restricted coloring problem (RCP)

In practical applications a conflict-free vertex/edge coloring is searched, often satisfying additional requirements. Therefore, a large number of particular coloring problems arised and has been investigated [12].

One well known example is vertex coloring with some restrictions set on available colors for the given graph vertex. In RCP each vertex is assigned a list of forbidden colors and a proper solution meeting such set of constraints is searched [13].

III. Sequential and parallel metaheuristics

The reported research is based on two sequential and nine parallel algorithms. The sequential metaheuristics include classical simulated annealing (SA) and tabu search (TS) that belong to the class of iterative methods [16]. Parallel algorithms can be splitted into three categories: parallel metaheuristics derived from SA, parallel metaheuristics derived from TS and hybrid methods.

A. Simulated annealing (SA)

Classical simulated annealing [16] is a well known technique widely used in optimization and present in most of the textbooks. It can be easily parallelized in various ways. Parallel moves enable single Markov chain to be evaluated by multiple processing units calculating possible moves from one state to another. Multiple threads compute independent chains of solutions and periodically exchange the obtained results. The key question in parallel implementation remains setting of algorithm's parameters like initial temperature, and a cooling schedule. For the problem at hand it is necessary to define an appropriate solution representation, cost function and a neighborhood generation scheme.

B. Tabu search (TS)

Tabu search [16] is an improvement of local search method in which so called tabu list contains a number of recent moves that must not be considered as candidates in the present iteration. This feature helps the method to escape from local minima what is impossible in local search. The question is to define the solution representation, cost function, neighborhood and a single move, the size of the neighborhood and the number of candidate moves, aspiration level which decides on the possibility to accept forbidden moves if it leads to a solution improvement etc.

C. MIR model of parallelization

Multiple independent runs (MIR) model is a very popular way of parallelization of iterative algorithms. A number of algorithm instances with different input data are executed simultaneously. All computational processes run independently and do not exchange data during computation. At the end, the best solution from all processes is selected. This simple model can be made more sophisticated by introducing an information exchange scheme, exchange rate etc.

D. MS model of parallelization

In Master-Slave (MS) model the master executes the sequential part of an algorithm, distributes computational tasks among slaves, collects results from slaves, process and aggregates this results. In certain versions of MS model the master splits the whole search space among slaves, synchronizes their work, checks the termination condition and collects the best solution from subspaces.

E. PA model of parallelization

Parallel asynchronous (PA) model provides maximum flexibility: various algorithms with different initial data search the whole search space in an asynchronous manner. Usually an efficient update scheme for the best solution must be implemented as well as occasional distribution of best solutions to asynchronous computational processes. One possibility is to employ a communication process. In some cases shared memory (SM) can be used for information updates and exchange. The second solution helps to avoid generation of interrupts in asynchronous processes. The processes communicate the SM in predictable moments of time.

F. Hybrid models

Hybrids models include : 1. two-phase algorithms, when each phase - restriction of the search space and solution refinement - is performed by a different method; 2. combined algorithms, when known elements of existing methods are composed in a single algorithm; 3. combined algorithms consisting original components like problem-oriented operations or heuristics; and 4. concurrent algorithm which is parallel execution of known methods with data exchange patterns.

In this paper three heuristic algorithms are used.

Parallel hybrid asynchronous (H-PA) algorithm splits computational processes into "even" performing SA and "odd"

performing TS. Best solutions are updated via shared memory SM, where are immediately made available for all processes.

Hybrid serial-parallel algorithm (H-SP) process in parallel p threads in which SA and TS sections are performed alternatively starting from SA section. SA section modifies tabu list while TS section modifies current temperature for the next section, respectively. Swiching conditions are related to the progress achieved in improving best solution.

Parallel hybrid algorithm (H-P) is developed on the basis MIR method. Single step combines properties of both SA and TS: if new solution satisfies aspiration criterion (AC) it is always accepted, otherwise, it is accepted according to SA rules. This means that probability of acceptance of worst solution decreases in time.

IV. THE SOLVER

For all tests it is used the "Partitioning problems solver" application. It is written in C++ (Visual Studio), while .NET Framework 3.5 provides necessary libraries and runtime environment.

The main program window contains three tabs: Program, Generator and Help. In appropriate fields of Program tab it is possible to select one of five basic problems (GPP, CPP, CPP-MIN, GCP, RGCP) and one of eleven algorithms. After that one can select the input file format and read input data. A numerous algorithm parameters and problem constrains must be filled in the forms including multiple runs, enabling statistics and write options. The cost of best solution and the total computation time are also displayed in this tab.

The Generator tab opens possibilities to generate input graphs or weighted input graphs after setting its parameters and lists of forbidden colors. The unweighted graphs are kept in .col format, weighted graphs are in .ecl format, which is extention of .col by adding edge weights as well as edge weight range (in the header). The type .rcp contains lists of forbidden colors for all vertices, if any. File formats .xpp and .xcp are used for preserving input graph and the partition being the best solution for the given problem together with its cost, respectively. Output data in CSV format are written to the .txt file and enable easy import of data into a spreadsheet.

V. COMPUTATIONAL EXPERIMENTS

For experiments the Intel Pentium T2300 machine was used with two 1,66 GHz cores and 4GB RAM, running under Windows XP Pro SP2 and .NET Framework 3.5 platform.

All five problems were tested agaist all eleven algorithms with eight basic settings (stop criterion, no of iterations in a single step, initial temperature for SA, size of the tabu list). The specific setting that were selected in the initial phase of the experiment are shown in Table I.

Other parameters are: coefficient of cost function = 1, no of parallel processes (if any) = 20, communication parameter = 20, no of algorithm repetitions = 20, no of clique extention trials = 5, no of repetitions for H-SP algorithm = 5.

In Tables II-XI computational data are presented. All experiments were conducted for random graph instances generated

TABLE I
BASIC SETTINGS OF ALGORITHMS

no.	stop criterion (it)	number of iterations/ step	SA - initial temperature	TS - size of tabu list
1	20	5	3	10
2	20	5	10	40
3	20	10	3	10
4	20	10	10	40
5	50	5	3	10
6	50	5	10	40
7	50	5	3	10
8	50	5	10	40

for each class of the graph partitioning problems in .ecl format. Relatively small graph instances were used with 20, 50 and 100 vertices and graph densities 10%, 20% and 30%. Cost functions from 20 trials are collected in Tables II-VI while the corresponding computation times in Tables VII-XI, respectively.

Average (Avg.) values for settings 1-8 in Tables II-XI are computed for parallel and hybrid methods only, serial methods SA i TS are excluded. Analysis of the results obtained for the five partitioning problems justifies several conclusions.

The shortest processing times are obtained by pure TS and SA methods. However, their solutions are not satisfactory. Parallelization and hybridization require additional computational work, and their aim is to improve search for a better suboptimal solution rather then providing significant speedup.

For GPP the fastest parallel algorithms are PTS metaheuristics. PSA and hybrid methods are less timeefficient. The slowest algorithm is H-SP, which is very time consuming. On the other hand H-SP finds the best solutions for all eight available settings. Average results of PSA-MIR and H-P algorithms are also outstanding and obtained approximately five times faster than by H-SP. The best setting in average is no 6 (minimum cost for six methods), but the best result for GPP is obtained with setting no 5. In terms of the computation time settings no 2 and 1 obviously win, and the fastest method is the PTS-A algorithm with moderate success in optimization.

For CPP the fastest parallel algorithms are PSA metaheuristics. Five other methods, except H-PS are also timeefficient. Among the parallel algorithm PSA-A is the fastest one with minimum time obtained for four settings. The slowest algorithm is again H-SP, which finds the best solutions for all eight parameter settings. The second result provides PTS-MIR which is eight times faster than H-SP. Setting no 8 provides the best solution quality for 7 parallel algorithms. In terms of the computation time settings no 2 and 1 win.

For CPP-MIN the fastest parallel algorithms is one hybrid and all PSA metaheuristics. The winner is PTS-MS algorithm with setting no 1. The slowest algorithm is H-SP, which wins the quality competition for all eight parameter settings. PSA-MIR and H-P have been the most prospective challengers. Setting no 8 provides the best solution quality for eight algorithms. In terms of the computation time settings no 2 and 1 are the winners.

For GCP the fastest parallel methods are H-PA and all PSAs which provide also best approximate solutions (PSA-MIR wins for six out of eight settings). The fastest parallel algorithm is H-PA, the best setting for six algorithms is no 2. The slowest algorithm is H-SP, which is 5th in terms of solution quality. The best setting for cost-optimality is no 4 in average.

The final problem - RCP - brings also interesting results. The fastest parallel algorithms are H-P six winning settings and PSAs. The best settings for all methods are 1 and 2. The best solution in average is found by PSA-MIR (the winner for seven out of eight settings), the runner-up is H-SP which was about eight time s lower, the next positions are occupied by PSA-MS and PSA-A. Most good results (8) were obtained for the setting no 8.

VI. Conclusions

In this paper some research results related to parallel metaheuristics and their applications were reported. The conducted experiments gave certain limited insight to computational behaviour of parallel metaheuristics developed on the basis of SA and TS, and applied to a class of popular partitioning problems in graphs. Some algorithms were better then others for solving particular problems. We were focused mostly on solution quality, but computation time was the second factor in comparison. Many results were not obvious and difficult to predict without verification. We believe that the presented initial results justify further experiments with our solver for more elaborated input instances. For this purpose to chose and modify DIMACS graph coloring instances, which were used for generation of instances of such partitioning problems like sum coloring, robust coloring etc. In time we will improve the algorithms to obtain more accurate solutions.

VII. Acknowledgements

The presented research was conducted within the frame of statutory activity (DS/2016) at Faculty of Electrical and Computer Engineering, Cracow University of Technology and was financially supported by Ministry of Science and Higher Education, Republic of Poland.

References

[1] G. Ausiello. et all. Complexity and Approximation - Combinatorial optimization problems and their approximability properties, Springer-Verlag, 1999.

[2] C. Blum, A. Roli, E. Alba. An Introduction to Metaheuristic Techniques. [in:] Parallel Metaheuristics, Wiley-Interscience, 3-42, 2005

[3] U. Brandes, M. Gaertler, D. Wagner Experiments on Graph Clustering Algorithms, Proc. ESA'2003, LNCS 2832, 568-579, 2003 DOI: 10.1007/978-3-540-39658-1-52

[4] C-Y. Byun. Lower Bound for Large-Scale Set Partitioning Problems, *ZIB-Report*, Vol. 12, 1822, 2001

[5] I. Charon, O. Hudry. Noising methods for a clique partitioning problem. *Discrete Applied Mathematics*, Vol. 154 , 754-769, 2006 DOI: 10.1016/j.dam.2005.05.029

[6] L. Coslovich, R. Pesenti, W. Ukovich. Large-Scale Set Partitioning Problems. *Journal on Computing*, Vol. 13, 191-209, 2001

[7] B. Crawford, C. Castro. ACO with Lookahead Procedures for Solving Set Partitioning Problems and Covering Problems. Chile, 2004

[8] R. Garey, D.S. Johnson. Computers and intractability. A guide to the theory of NP-completeness. Freeman, San Francisco, 1979

[9] X. Ji, J.E. Mitchell. The Clique Partition Problem with Minimum Clique Size Requirement, *Discrete Optimization*, Vol. 4, 87-102, 2007

[10] B.W. Kernighan, S. Lin. An Efficient Heuristics Procedure for Partitioning Graphs. *The Bell System Technical Journal*, Vol. 49, 291-307, 1970

[11] Z. Kokosiński, K. Kwarciany, M. Kołodziej. Efficient Graph Coloring with Parallel Genetic Algorithms. *Computing and Informatics*, Vol. 24, No. 2, 109-121, 2005

[12] M. Kubale(Ed.) Graph Colorings, American Mathematical Society, 2004

[13] M. Kubale. Some results concerning the complexity of restricted colorings of graphs. *Discrete Applied Mathematics*, Vol. 36, 35-46, 1992

[14] E. Mujuni, F. Rosamond. Parameterized Complexity of the Clique Partition Problem. *The Australasian Theory Symposium*, 75-78, 2008

[15] A. Nowosielski, P.A. Kowalski, P. Kulczycki. The column-oriented database partitioning optimization based on the natural computing algorithms Proc. FedCSIS, 1035-1041, 2015 DOI : 10.15439/2015F262

[16] S.M. Sait, H. Youssef. Iterative computer algorithms with applications in engineering. IEEE Computer Society, Los Alamitos, 1999

[17] W-D. Tseng, I-S. Hwang, L-J. Lee, C-Z. Yang. Clique-partitioning connections-scheduling with faulty switches in dilated Benes network, *Journal of the Chinese Institute of Engineers*, Vol. 32, 853-860, 2009

[18] S. Zhou. Minimum partition of an independence system into independent sets, *Discrete Optimization*, Vol. 6, 125-133, 2009 DOI: 10.1016/j.disopt.2008.10.001

TABLE II
GRAPH PARTITIONING PROBLEM (GPP). COST FUNCTIONS (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	295	238	255	251	338	284	285	287	286	230	240	262
2	296	237	254	253	330	286	285	282	246	228	234	256
3	293	243	257	259	338	283	285	288	254	234	238	260
4	293	238	257	257	331	282	285	283	252	232	235	260
5	292	234	249	247	341	276	286	288	245	**226**	237	254
6	286	235	245	244	338	274	284	281	238	227	235	**251**
7	289	242	252	256	332	280	286	286	249	238	242	259
8	289	240	252	252	329	271	280	283	252	235	239	256
Avg.	292	238	253	252	335	280	285	285	253	**231**	238	

TABLE III
CLIQUE PARTITIONING PROBLEM (CPP). COST FUNCTIONS (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	26	24	24	24	25	23	24	24	24	**21**	24	23,6
2	26	24	24	24	25	23	24	24	24	**21**	24	23,6
3	23	22	22	22	24	22	23	23	23	**21**	22	22,2
4	24	22	22	22	24	22	23	23	23	**21**	22	22,2
5	25	24	24	24	24	22	23	23	23	**21**	24	23,1
6	26	24	24	24	24	22	23	23	23	**21**	24	23,1
7	23	22	22	22	23	22	22	22	22	**21**	22	**21,9**
8	24	22	22	22	23	22	22	22	22	**21**	22	**21,9**
Avg.	24,6	23	23	23	24	22,3	23	23	23	**21**	23	

TABLE IV
CPP WITH MIN. CLIQUE SIZE (CPP-MIN). COST FUNCTIONS $x10^3$ (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	114	107	109	108	147	127	136	134	134	102	106	118
2	115	108	109	109	145	127	136	134	111	102	107	116
3	112	105	106	105	138	123	129	130	108	100	105	112
4	113	106	106	106	137	124	129	129	108	101	106	113
5	113	106	106	106	140	114	131	130	109	102	106	112
6	114	106	106	107	140	114	130	130	109	102	106	112
7	113	105	105	105	134	114	127	127	108	**99,3**	105	111
8	112	106	105	105	133	113	126	126	107	**99,3**	105	**110**
Avg.	113	106	107	106	139	119	131	130	112	**101**	106	

TABLE V
GRAPH COLORING PROBLEM (GCP). COST FUNCTIONS (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	86	51	52	53	84	55	55	55	55	55	54	53,9
2	80	52	52	52	85	55	54	55	53	53	53	53,2
3	85	52	51	52	83	54	55	54	54	54	53	53,2
4	86	51	52	51	78	53	53	53	53	53	53	**52,4**
5	82	52	52	52	86	54	56	55	54	54	53	53,6
6	87	52	52	52	86	53	54	55	54	53	53	53,1
7	84	**50**	51	52	83	53	54	55	54	53	53	52,8
8	86	52	51	52	85	53	54	53	52	53	53	52,6
Avg.	84,5	**51,5**	51,6	52	83,8	53,8	54,4	54,4	53,6	53,5	53,1	

TABLE VI
RESTRICTED GCP (RGCP). COST FUNCTIONS (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	36	**26**	**26**	27	39	29	30	29	30	27	29	28,1
2	37	**26**	**26**	**26**	37	28	29	29	27	26	28	27,2
3	36	**26**	**26**	27	39	28	29	28	28	**26**	28	27,3
4	36	27	27	27	36	28	28	28	27	**26**	28	27,3
5	37	**26**	**26**	**26**	38	28	29	29	27	**26**	28	27,2
6	36	**26**	27	**26**	38	27	28	29	27	27	28	27,2
7	36	**26**	27	27	37	28	28	28	27	**26**	28	27,2
8	36	**26**	27	27	36	27	28	28	27	**26**	27	**27,0**
Avg.	36,3	**26,1**	26,5	26,6	37,5	27,9	28,6	28,5	27,5	26,3	28	

TABLE VII
GRAPH PARTITIONING PROBLEM (GPP). COMPUTATION TIMES (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	1,70	19,4	14,6	15,3	0,57	6,83	5,23	5,13	5,49	70,0	34,2	19,6
2	1,76	18,3	13,4	14,2	0,60	6,75	5,40	5,51	14,1	67,5	18,4	**18,2**
3	6,28	64,2	48,4	49,8	2,48	26,5	21,9	21,3	40,1	27,0	64,9	67,5
4	5,71	58,3	42,8	41,5	2,85	29,8	26,9	25,6	40,1	258	58,7	64,7
5	3,77	35,0	32,2	30,4	1,52	25,3	13,1	13,5	27,8	149	35,4	40,1
6	3,33	34,4	31,3	32,4	1,76	33,4	15,1	15,0	28,2	149	34,9	41,5
7	10,7	106	97,7	98,1	6,59	114	55,4	54,8	85,9	626	109	150
8	10,2	101	91,7	92,5	6,98	128	67,1	65,6	86,5	605	104	149
Avg.	5,44	54,8	46,5	46,8	2,92	46,3	26,3	**25,8**	41,0	274	57,5	

TABLE VIII
CLIQUE PARTITIONING PROBLEM (CPP). COMPUTATION TIMES (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	1,03	9,86	10,0	9,82	1,05	10,8	10,0	10,5	10,5	76,4	10,0	17,5
2	0,95	9,46	9,59	9,40	1,06	11,0	10,5	10,4	10,2	76,6	9,62	**17,4**
3	3,45	33,9	33,3	32,7	3,30	34,0	31,8	31,5	32,7	280	33,7	60,4
4	3,09	30,9	30,5	30,8	3,33	33,3	31,6	31,9	32,6	282	31,1	59,4
5	1,89	18,2	18,3	18,4	2,14	23,5	21,5	21,7	22,1	168	18,2	36,7
6	1,78	17,6	17,9	17,9	2,27	23,5	22,0	21,6	21,2	158	17,4	35,2
7	6,50	63,8	63,8	63,5	7,25	77,9	70,2	71,4	73,5	653	64,8	134
8	6,10	61,8	60,9	60,9	7,30	76,3	72,5	70,2	72,4	645	61,2	131
Avg.	3,10	30,7	30,5	**30,4**	3,46	36,3	33,8	33,7	34,4	292	30,8	

TABLE IX
CPP WITH MIN. CLIQUE SIZE (CPP-MIN). COMPUTATION TIMES (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	1,74	17,6	16,3	17,2	1,25	18,1	13,0	13,6	13,3	91,2	18,0	24,3
2	1,76	17,0	16,7	17,3	1,28	17,5	13,4	13,6	16,6	89,4	16,6	**24,2**
3	3,89	39,9	39,6	39,4	3,82	53,3	40,5	39,3	40,0	337	39,7	74,4
4	3,90	38,4	38,2	39,2	3,90	52,1	39,4	39,5	40,1	332	38,4	73,1
5	2,93	27,6	29,6	29,3	2,60	50,5	26,7	26,4	30,0	193	26,9	48,9
6	2,85	27,4	29,9	29,3	2,58	49,6	27,1	27,3	29,7	185	27,4	48,1
7	7,24	72,3	72,1	73,4	8,60	169	89,1	83,0	77,6	795	71,9	167
8	7,33	71,9	73,1	71,8	9,04	167	90,4	86,6	76,6	840	71,9	172
Avg.	3,96	39,0	39,4	39,6	41,4	72,2	42,4	41,2	40,5	358	**38,9**	

TABLE X
GRAPH COLORING PROBLEM (GCP). COMPUTATION TIMES (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	0,62	7,09	7,14	6,97	0,64	7,06	7,01	6,88	6,84	49,3	8,13	**11,8**
2	0,60	6,56	6,57	6,63	0,68	8,05	7,68	7,35	6,36	48,9	8,08	**11,8**
3	2,36	26,9	26,2	26,2	2,38	27,2	24,6	25,2	23,3	201	27,9	45,4
4	2,24	25,1	24,6	24,9	2,31	27,1	26,0	26,1	22,4	201	27,5	44,9
5	1,37	14,4	14,4	14,3	1,51	17,9	15,9	15,6	15,6	121	18,3	27,5
6	1,31	13,9	14,0	14,0	1,45	19,2	17,6	18,0	20,4	122	18,9	28,7
7	5,35	56,9	56,4	56,1	5,73	65,1	61,3	60,0	55,8	498	68,3	109
8	5,24	54,5	55,3	54,7	5,37	61,3	56,3	56,8	52,8	499	59,6	106
Avg.	2,39	25,7	25,6	25,5	2,51	29,1	27,1	27,0	**25,4**	218	29,6	

TABLE XI
RESTRICTED GCP (RGCP). COMPUTATION TIMES (11 ALGORITHMS, 8 SETTINGS, 20 RUNS)

	SA	**PSA**			TS	**PTS**			**Hybrid**			**Avg.**
		MIR	MS	A		MIR	MS	A	H-PA	H-SP	H-P	
1	0,77	7,73	7,58	7,92	0,66	7,58	6,56	6,61	6,59	49,0	7,35	**11,9**
2	0,72	7,15	7,25	7,07	0,71	7,80	7,35	7,24	6,20	49,1	8,01	**11,9**
3	2,78	28,7	28,9	29,4	2,51	27,1	25,7	25,2	23,1	199	28,1	46,2
4	2,64	26,9	26,7	26,7	2,57	28,7	26,9	26,2	22,9	199	28,9	45,9
5	1,51	14,9	15,2	15,3	1,58	18,3	15,6	16,1	14,1	121	18,4	27,7
6	1,45	14,3	14,5	14,6	1,68	19,2	17,5	16,5	14,6	121	19,2	28,0
7	5,84	57,9	58,6	58,7	6,26	72,3	63,6	62,7	55,0	497	72,4	111
8	5,61	56,4	56,1	56,7	5,79	66,0	63,6	61,9	55,3	496	66,6	109
Avg.	2,66	26,8	26,9	27,1	2,72	30,9	28,3	27,8	**24,7**	217	31,1	

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 87–93
DOI: 10.15439/2016F554
ACSIS, Vol. 9. ISSN 2300-5963

Is Your Parallel Algorithm Correct?

Jakub Nalepa
Institute of Informatics
Silesian University of Technology
Akademicka 16
44-100 Gliwice, Poland
Email: jakub.nalepa@polsl.pl

Miroslaw Blocho
Institute of Informatics
Silesian University of Technology
Akademicka 16
44-100 Gliwice, Poland
Email: blochom@gmail.com

***Abstract*—Verifying the correctness of parallel algorithms is not trivial, and it is usually omitted in the works from the parallel computation field. In this paper, we discuss in detail how to show that a certain parallel algorithm is correct. This process involves proving its safety and liveness. We perform the in-depth analysis of our parallel guided ejection search (P–GES) for the pickup and delivery problem with time windows, which serves as an excellent case study. P–GES was implemented as a distributed algorithm using the Message Passing Interface library with asynchronous communications, and was validated using the well-known Li and Lim's benchmark containing demanding test instances. We already proved the efficacy of this algorithm and showed that it can retrieve very high-quality (quite often better than the world's best at that time) routing schedules.**

I. Introduction

DESIGNING and implementing parallel algorithms attracted attention of researchers from various fields, including the computational biology, genomics, text processing, pattern recognition, machine learning, optimization, and many others, due to the availability of various parallel architectures. Such approaches allow for solving extremely complex tasks in short time, assuming that: the parallel algorithms are correct and scalable. Proving the correctness of parallel techniques is not trivial (it is much more difficult compared with serial algorithms), and it is omitted in a majority of works belonging to the parallel computation field.

In this paper, we show how to investigate the correctness of a given parallel algorithm. Our parallel guided ejection search technique (P–GES) for minimizing the number of trucks in the NP-hard pickup and delivery problem with time windows, serves as the case study—we analyze its correctness, and show how to accomplish that in a step-by-step manner. In our previous works [1], [2], we experimentally evaluated the Message Passing Interface implementation of P–GES. The extensive experimental study revealed that this algorithm is quite efficient, and it is able to extract very high-quality feasible schedules (often better than the world's best known solutions at that time). The analysis of its correctness presented in this paper therefore complements our previous efforts and theoretically proves that P–GES is correct indeed.

This paper is structured as follows. Section II gives the formulation of the pickup and delivery problem with time windows (PDPTW). Section III reviews the state of the art on solving the PDPTW, and on parallel heuristic algorithms, in order to better contextualize our parallel guided ejection search within the literature. In Section IV, we present the background on verifying the correctness of parallel algorithms, and the correctness of our parallel guided search is proven in Section V. The paper is concluded in Section VI, which also serves as the outlook to our future work.

II. Pickup and Delivery with Time Windows

The PDPTW is a problem of serving a number of transportation requests, each being a pair of the pickup and delivery requests. The PDPTW is therefore defined on a directed graph $G = (V, E)$, with a set V of $C + 1$ vertices. The vertices v_i, $i \in \{1, ..., C\}$, represent the travel points, whereas v_0 denotes the depot (the start and the finish point of each route). A set of edges $E = \{(v_i, v_{i+1}) | v_i, v_{i+1} \in V, v_i \neq v_{i+1}\}$ are the travel connections between each pair of travel points. The travel costs $c_{i,j}$, $i, j \in \{0, 1, ..., C\}$, $i \neq j$, are equal to the distances (in the Euclidean metric) between the travel points. Each request h_i, $i \in \{0, 1, ..., N\}$, where $N = C/2$, is a coupled pair of pickup (P) and delivery (D) customers—these customers are given as p_h and d_h, respectively, where $P \cap D = \emptyset$, and $P \cup D = V \setminus \{v_0\}$ (a customer cannot request both delivery and pickup operations). For each request h_i, the amount of delivered ($q^d(h_i)$) and picked up ($q^p(h_i)$) goods is defined, where $q^d(h_i) = -q^p(h_i)$. Hence, each customer v_i defines its own demand (this is either the delivery or the pickup demand), service time s_i (note that "serving" the depot does not take time, and $s_0 = 0$), and time window $[e_i, l_i]$ within which the service of this customer should be *started* (however, it can be finished after closing this time slot). Since the fleet is homogenous (let K denote its size), the capacity of each truck is equal (it is given as Q). Each route r, given as $r = \langle v_0, v_1, ..., v_{n+1} \rangle$ in the solution σ (being a set of routes), starts and finishes at the depot, thus $v_0 = v_{n+1}$, and it is an ordered list of visited travel points.

An exemplary PDPTW solution (σ) is rendered in Fig. 1—22 customers (they are divided into the pickup and delivery ones, hence there are 11 pickup-delivery requests) are served in the following three routes: $r_1 = \langle v_0, v_6, v_2, v_1, v_3, v_5, v_4, v_0 \rangle$ (3 requests are handled), $r_2 = \langle v_0, v_8, v_{10}, v_{13}, v_{11}, v_{12}, v_{17}, v_0 \rangle$ (3 requests), $r_3 = \langle v_0, v_{14}, v_{15}, v_{19}, v_{22}, v_{21}, v_{20}, v_{18}, v_{16}, v_9, v_7, v_0 \rangle$ (5 requests). Assuming that (i) the vehicle capacity Q is not

exceeded for any vehicle (capacity constraint is satisfied), (ii) the service of every customer starts within its time window (time window constraint), (iii) every customer is served in exactly one route, (iv) every vehicle starts at and returns to the depot within the time window of the depot ($[e_0, l_0]$), and (v) each pickup is performed before the corresponding delivery for each request (precedence constraint), then this solution is feasible.

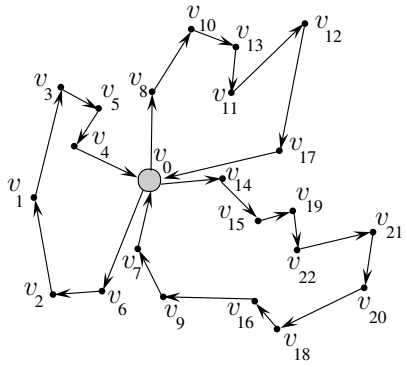


Fig. 1. Exemplary PDPTW solution: 22 clients (11 requests) served in 3 routes.

The PDPTW is a two-objective NP-hard discrete optimization problem. Its primary objective is to minimize the fleet size K, whereas the second objective is to minimize the distance $T = \sum_{i=1}^{K} T_i$, where T_i is the distance traveled in the i-th route. Let σ_A and σ_B denote two feasible PDPTW solutions. σ_A is then of a higher quality compared with σ_B, if $(K(\sigma_A) < K(\sigma_B))$ or $(K(\sigma_A) = K(\sigma_B)$ and $T(\sigma_A) < T(\sigma_B))$. Hence, the solution is of a higher quality if it consists of a lower number of routes, or—if the number of trucks is equal for both solutions—if the total travel distance is shorter.

III. Related Literature

A. Solving the Pickup and Delivery with Time Windows

State-of-the-art algorithms for rich routing problems encompass exact and approximate methods [3]. The former algorithms deliver the exact solutions [4], [5], [6], [7], however they are very difficult to apply in practice, because of their unacceptably large execution times (especially in the case of massively large, real-life problem instances). Also, handling the dynamic changes which are very common in many circumstances (e.g., updating the traffic networks to avoid congestion) are not trivial to incorporate in such algorithms [8]. The exact techniques were discussed in several works [9], [10].

Approximation algorithms include construction and improvement heuristics and metaheuristics [11], [12]. The construction (insertion-based) techniques create solutions from scratch by inserting consecutive requests iteratively into the partial solution [13], [14]. The partial solution encompasses a subset of all transportation requests, therefore is not acceptable and should be expanded to serve other requests (feasibly) as well. On the other hand, improvement heuristics modify an initial solution (very often of a low quality) by applying local search moves (thus, by exploring the neighborhood of this solution) [15], [16]. A number of metaheuristics have been adopted for solving rich VRPs throughout the years, including various tabu searches [4], variable neighborhood searches [17], greedy randomized adaptive search procedures, population-based [18], [19], [20], and agent-based approaches [21], guided ejection searches [22], simulated annealing [16], and more [23].

B. Parallel Heuristic Algorithms

Parallel heuristic algorithms have been explored for solving a bunch of different optimization problems [24], including various VRPs [21], [25]. Co-operative strategies in such parallel heuristic techniques have been discussed and classified in several taxonomies, with the one presented by Crainic et al. being the best established [26], which encompasses three dimensions. The first dimension specifies if the global solving procedure is controlled by a single process (1-control—1C) or by a group of processes (p-control—pC). These processes may co-operate (in co-operative algorithms) or not (if the processing is batched). The second dimension reflects the quantity and quality of the information exchanged between the parallel processes, along with the additional knowledge derived from these exchanges. The four classes are defined for this dimension: Rigid (RS), Knowledge Synchronization (KS), Collegial (C) and Knowledge Collegial (KC). The third dimension concerns the diversity of the initial solutions and search strategies: Same Initial Point / Population, Same Search Strategy (SPSS), Same Initial Point / Population, Different Search Strategies (SPDS), Multiple Initial Points / Populations, Same Search Strategies (MPSS), Multiple Initial Points / Populations, Different Search Strategies (MPDS).

The parallel algorithms were very intensively explored for solving rich routing problems [25], [27], including the PDPTW [1], [2]. The implementations of these algorithms take advantage from the massively-parallel architectures (both with the shared and distributed memory [28]) which are easily accessible nowadays. These techniques deliver extremely high-quality routing schedules in short time, even for enormously large problem instances.

IV. Verification of the Correctness of Parallel Algorithms

Verifying the correctness of a given sequential algorithm encompasses showing that this algorithm: (i) will finish (thus, the termination conditions will be finally met), and (ii) will give a correct result for any correct set of input data. More formally, the correctness may be stated as:

$$\{p\}A\{q\}, \tag{1}$$

where A denotes the algorithm (a set of statements), p is the pre-condition, and q represents the post-condition. Here, the pre-condition specifies which conditions must hold for the input data, and the post-condition reflects what should be

satisfied by the results retrieved using the algorithm A. The algorithm is *partially correct* if for any input data satisfying the pre-condition, it gives the correct output data (in accordance with the post-condition) [29] (thus, the input-output relation holds). On the other hand, the algorithm is *totally correct*, if it is partially correct, and—for any input data—it reaches the termination condition (this is not crucial in the case of the partial correctness), and returns the correct output. It is easy to note that proving the total correctness of a sequential algorithm may consist of proving its partial correctness, along with showing that every execution of this algorithm will result in meeting the stopping condition [30].

In the case of parallel algorithms, proving their correctness includes verifying their *safety* and *liveness*. The algorithm is safe, if it can never end up in a forbidden state. To prove that, we need to show that the algorithm is (i) partially correct, (ii) there are no deadlocks (i.e., the processes do not wait for the infinite amount of time for each other to continue the execution), and that (iii) only the processes can safely access the shared resource (*mutual exclusion*). The liveness property of a parallel algorithm is satisfied, if it can be proven that a certain desired condition will eventually happen during the execution of this algorithm [31], [32]. In the case of message-passing techniques—as shown in [30]—it is important to show that the messages are properly sent and received (no matter if the communication is synchronous or asynchronous).

If all of the above-mentioned properties of an analyzed parallel algorithm are proven, then this algorithm is *correct*.

V. Correctness of the Parallel Guided Ejection Search for the PDPTW

The baseline (sequential) version of the GES was proposed in [23], and later enhanced and parallelized in our very recent works [1], [2], [22]. According to the taxonomy mentioned in Section III-B, P–GES is of the pC/C/MPSS type (p-Control, Collegial, Multiple Initial Points, Same Search Strategies).

A. Algorithm Outline

In P–GES, which is an improvement parallel heuristic technique, *p* processes execute in parallel (Algorithm 1, line 1). The initial feasible solution σ contains the number of routes which is equal to the number of transportation requests, hence each request is feasibly served by a separate truck (line 3). Then, the number of serving vehicles in σ is consecutively decreased until the total computation time exceeds the imposed time limit τ_M (lines 5-36), or the desired number of routes has been obtained.

A random route *r* is removed from σ, and the excluded requests are put into the ejection pool (EP), which stores those transportation requests that have been removed from the schedule—this solution becomes partial (line 7). The penalty counters (indicated as p's), which reflect the difficulty of re-inserting a given request back into the partial solution, for all of the requests are reset (line 8).

If the EP contains unserved transportation requests (lines 9-32), then a single request h_{in} is popped from the EP at the time (line 10), and it is being re-inserted into the partial solution. If there exist any feasible insertion positions for this request (the set of such positions $S_{in}^{fe}(h_{in}, \sigma)$ is not empty), then a random position is drawn (line 12). If it is not the case, then the request is inserted into σ infeasibly (so that it violates the constraints), and the feasibility of the (possibly partial) solution is being restored in the *squeezing* procedure (line 14). Here, the solution penalty is quantified using the penalty function given as:

$$\mathcal{F}_p(\sigma) = \mathcal{F}_c(\sigma) + \mathcal{F}_{tw}(\sigma), \tag{2}$$

where $\mathcal{F}_c(\sigma)$ and $\mathcal{F}_{tw}(\sigma)$ are the sum of capacity exceeds in σ, and the sum of the time windows violations, respectively. The squeeze function (presented in Algorithm 2) aims at

Algorithm 1 A parallel algorithm to minimize K (P–GES).

```
1: for P_i ← P_1 to P_p do in parallel
2:     τ_last ← τ_curr;
3:     Create an initial solution σ;
4:     finished ← false;
5:     while not finished do
6:         Save the current feasible solution;
7:         Put requests from a random route r into EP;
8:         Set penalty counters p[i] ← 1 (i = 1, 2, ..., N);
9:         while (EP ≠ ∅) and (not finished) do
10:            Select and remove request h_in from EP;
11:            if S_in^fe(h_in, σ) ≠ ∅ then
12:                σ ← random σ' ∈ S_in^fe(h_in, σ);
13:            else
14:                σ ← Squeeze(h_in, σ);
15:            end if
16:            if h_in is not inserted into σ then
17:                p[h_in] ← p[h_in] + 1;
18:                for k ← 1 to k_m do
19:                    Get S_ej^fe(h_in, σ) with min. P_sum;
20:                    if S_ej^fe(h_in, σ) ≠ ∅ then
21:                        σ ← random σ' ∈ S_ej^fe(h_in, σ);
22:                        Add (h_out^(1), h_out^(2), ..., h_out^(k)) to EP;
23:                        break;
24:                    end if
25:                end for
26:            end if
27:            σ ← Perturb(σ);
28:            if τ_curr ≥ τ_last + τ_coop then
29:                finished ← Cooperate(σ);
30:                τ_last ← τ_curr;
31:            end if
32:        end while
33:        if EP ≠ ∅ then
34:            Backtrack to previous feasible solution;
35:        end if
36:    end while
37:    Get the best solution σ_best;
38: end for
```

decreasing the value of this function until it reaches zero (thus, the solution is feasible). This is a steepest-descent local search procedure, in which the set $S^{inf}(h_{in}, r, \sigma_t)$ of infeasible solutions is created (considering the insertion of the analyzed transportation request), and the solution with the minimum value of the penalty function is picked up. This process continues until the feasibility is restored, or it is impossible to retrieve a feasible solution (in this case, the solution is backtracked to the initial state).

Algorithm 2 Squeezing an infeasible (possibly partial) solution σ.

```
1: function SQUEEZE(h_in, σ)
2:     σ_t ← σ' ∈ S^inf(h_in, σ) such that F_p(σ') is minimum;
3:     while (F_p(σ_t) ≠ 0) do
4:         Randomly choose an infeasible route r in σ_t;
5:         Find σ'' ∈ S^inf(h_in, r, σ_t) with min. F_p(σ'');
6:         if F_p(σ'') < F_p(σ_t) then
7:             σ_t ← σ'';
8:         else
9:             break;
10:        end if
11:    end while
12:    if F_p(σ_t) = 0 then
13:        return σ_t;
14:    else
15:        return σ;
16:    end if
17: end function
```

If the squeeze fails (thus the solution has been backtracked to the previous partial schedule), the penalty counter of the appropriate request ($p[h_{in}]$) is increased (Algorithm 1, line 17), and other requests are ejected from the solution (up to k_m requests; lines 18-25) to insert h_{in} (this request is of a "high priority"). The set $S^{fe}_{ej}(h_{in}, \sigma)$ is formed, and it encompasses the solutions with various combinations of ejected requests (the h_{in} request is inserted to this solution on various positions). Finally, the solution σ'—with the minimum sum of the penalty counters is selected from $S^{fe}_{ej}(h_{in}, \sigma)$ (line 21). Clearly, the ejected requests are pushed to the EP, and should be re-inserted into σ later (line 22). The solution σ is finally perturbed by the local search procedures, in which I feasible (i.e., not violating the constraints) local moves (out-relocate and out-exchange) are executed for the search diversification (line 27). This procedure is visualized in Algorithm 3.

The parallel processes in P–GES co-operate periodically every τ_{coop} seconds (Algorithm 1, line 29) using the asynchronous co-operation scheme. In our previous works [25], [2], we investigated a number of co-operation schemes (they define the co-operation topology, frequency, and the strategies for handling emigrants/immigrants) and showed, that a proper selection of such scheme has a tremendous impact on the algorithm capabilities and behavior.

In P–GES, it is the master process (P_1) which controls the execution time of the algorithm—the signals from P_1 to either continue or stop the execution are transferred in each co-operation phase. Eventually, all solutions from all processes are gathered in the master, and the best solution σ_{best} is retrieved—this is the final solution delivered by P–GES (line 37).

Algorithm 3 Perturbing a feasible (possibly partial) solution σ for the search diversification.

```
1: function PERTURB(σ)
2:     σ_t ← σ;
3:     for i ← 1 do I
4:         Find σ' through local search moves on σ_t;
5:         if σ' is feasible then
6:             σ_t ← σ';
7:         end if
8:     end for
9:     return σ_t;
10: end function
```

More details on P–GES can be found in our previous works [1], [2]. These papers include the in-depth analysis of the Message Passing Interface implementation of this algorithm, and discuss the experimental results retrieved for very demanding Li and Lim's benchmark sets (encompassing tests of various sizes and characteristics, e.g., positions of the travel points, and tightness of time windows).

B. Proving the Correctness of P–GES

The input data passed to P–GES include:

- p ($p \geq 1$)—the number of parallel processes. If $p = 1$, then P–GES becomes a sequential algorithm, and its certain components are disabled (e.g., the co-operation between the processes).
- $K_d \geq 0$—the desired number of trucks serving the requests. If $K_d = 0$, then the best feasible solution found using P–GES is returned (i.e., there is no "desired" number of routes, however K should be as minimum as possible).
- τ_{MAX}—the maximum execution time (in seconds) of P–GES.
- τ_{coop}—the co-operation frequency (in seconds).
- k_m ($k_m \geq 1$)—the maximum number of requests that can be ejected from a (possibly partial) solution while inserting a request popped from the EP.
- I ($I \geq 0$)—the number of local search moves applied to perturb a solution.
- Test instance—the definition of the test instance at hand. It specifies the number of transportation requests, the positions of the travel points, their time windows, service times, and demands (either pickup or delivery), and the maximum capacity of trucks. It is worth noting that real-life problems may encompass travel points which are clustered, randomly scattered around the map, or combine both (i.e., there are some customer clusters, but lots of them are random). The problem instances belonging to the Li and Lim's benchmark set perfectly reflect these

scenarios—the exemplary instance structures (with 100 travel points) are visualized in Table I.

The desired solution retrieved using P–GES must satisfy all the constraints discussed in Section II. Therefore, this solution must be *feasible* (otherwise, the routing schedule is incorrect).

As mentioned in Section V-A, P–GES starts with an initial feasible solution (therefore, the constraints are not violated), in which every transportation request is served in a separate route. Then, the attempts to reduce the fleet size are undertaken, until the execution reaches the termination condition (Algorithm 1, line 5)—one random route is analyzed at any time. The current (best) solution is saved (line 6). If removing this route fails, then the partial solution is backtracked to this state (line 34), hence the schedule remains feasible.

Once the ejected customers are pushed into the EP, the solution becomes a partial feasible schedule (no constraints are violated). Then, these transportation requests are put back into the partial solution—first, using the feasible insertion positions (if any). In this case, the feasibility is not violated, and the next request from the EP is popped for insertion. On the other hand, the infeasible solution (in which the request has been re-inserted back infeasibly) is processed with the squeeze procedure. This squeezing retrieves either the feasible solution (if it is possible to restore the correctness of σ using local search moves), or backtracks to the state before this squeezing has been called. In the latter case, other transportation requests are ejected to restore the feasibility of the partial solution, thus it finally becomes feasible. Perturbing a feasible (potentially partial) schedule can deteriorate its quality, however it cannot cause violating the constraints—after calling this procedure, the solution remains feasible. If the EP is empty, then the feasible solution—with the decreased fleet size—becomes the next solution, which is to be processed in the next algorithm iteration. Therefore, the PDPTW solution obtained using P–GES is eventually always feasible.

The co-operation of parallel processes cannot affect the feasibility of the solutions—depending on the co-operation scheme, the receiving process may e.g., replace its own solution with the immigrant (if the immigrant is of a higher quality). Clearly, this operation cannot affect the feasibility of the considered routing schedule. This analysis shows that P–GES is *partially correct*—assuming that the input data are correct, it always retrieves feasible PDPTW solutions.

P–GES may be terminated if either a solution of a desired quality (i.e., with the desired number of routes, K_d) is found, or if the maximum execution time elapsed. In the former case, this solution may be retrieved by the master process (which also controls the execution of other processes in the team, and may send the termination request), or any other process. If the master got this solution, then it sends the termination requests to others (thus, one co-operation phase is enough to stop the parallel algorithm execution). However, if another process ended up with the desired solution, then two co-operation phases would be necessary—first, it sends its best solution to the master, and then the master sends the termination request to other processes. In either case, P–GES finally reaches its stopping condition.

P–GES is a distributed algorithm (there are no shared resources). The co-operation is asynchronous (independently from the selected scheme), and the execution (i.e., optimization of the solution run by a given process) interleaves with the send/receive operations. The order of send/receive operations matter in this case, thus they are executed in an appropriate order depending on the process type (either the master on non-master). Additionally, receiving data is acknowledged by the receiving process during the co-operation (the status of this acknowledgement is periodically checked by the sending process). Since there are no deadlocks and shared resources in P–GES, its safety is proven. The same reasoning may be used to prove the liveness of the algorithm. Since only the master process can force other processes to stop, the situation in which a given process sends to or waits for a message from the process that has already been terminated is not possible. This shows the liveness property of P–GES.

The above investigation revealed that all of the conditions imposed on the parallel algorithms which ensured that the corresponding algorithm is correct are fulfilled by P–GES, for the correct input data (e.g., assuming that the test instance at hand is solvable). Therefore, P–GES is a correct parallel algorithm.

□

VI. Conclusions and Outlook

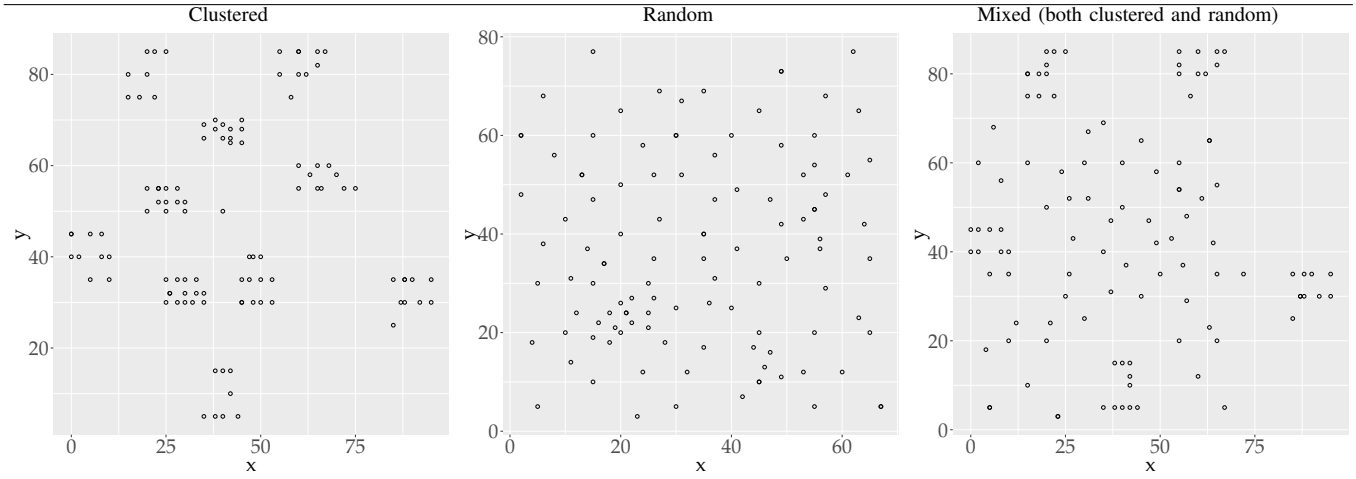
In this paper, we analyzed the correctness of our parallel guided ejection search algorithm for solving the PDPTW. We proved that the algorithm is correct—this involved showing its liveness and safety. This investigation served as an extensive case study for showing how to prove the correctness of parallel algorithms. This approach may be easily tailored for proving the correctness of other parallel algorithms, especially those tackling complex (however not only transportation) discrete optimization problems.

Our current research is focused on implementing a parallel memetic algorithm (a hybrid of a genetic algorithm and some local refinement procedures) for minimizing the travel distance in the PDPTW. Memetic algorithms were proven extremely efficient in solving a wide range of optimization and pattern recognition problems [33], [34], [35], [36], [37], [38]. Then, we will work on a parallel version of this algorithm (we already proved that our parallel memetic approach for the VRP with time windows is correct [30]). Combining the parallel guided ejection search discussed in this paper with the parallel memetic algorithm will enable us to create a full optimization framework for solving rich routing problems [39], especially the PDPTW.

Acknowledgment

This research was supported by the National Science Centre under research Grant No. DEC-2013/09/N/ST6/03461, and performed using the infrastructure supported by the POIG.02.03.01-24-099/13 grant: “GeCONiI—Upper Silesian

TABLE I
EXEMPLARY LI AND LIM'S INSTANCE STRUCTURES COMPOSED OF CLUSTERED, RANDOMIZED, AND MIXED CUSTOMERS (FOR 100 TRAVEL POINTS, BEING EITHER THE PICKUP OR DELIVERY CUSTOMERS).



Center for Computational Science and Engineering", and the Intel CPU and Xeon Phi platforms provided by the MICLAB project No. POIG.02.03.00.24-093/13.

REFERENCES

[1] M. Blocho and J. Nalepa, "A parallel algorithm for minimizing the fleet size in the pickup and delivery problem with time windows," in *Proc. of 22nd European MPI Users' Group Meeting*, ser. EuroMPI '15. New York, USA: ACM, 2015, pp. 15:1–15:2. [Online]. Available: http://doi.acm.org/10.1145/2802658.2802673

[2] J. Nalepa and M. Blocho, "A parallel algorithm with the search space partition for the pickup and delivery with time windows," in *10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2015, Krakow, Poland, November 4-6, 2015*, 2015, pp. 92–99. [Online]. Available: http://dx.doi.org/10.1109/3PGCIC.2015.12

[3] L. Grandinetti, F. Guerriero, F. Pezzella, and O. Pisacane, "The multi-objective multi-vehicle pickup and delivery problem with time windows," *Social and Beh. Sc.*, vol. 111, pp. 203 – 212, 2014.

[4] W. P. Nanry and J. W. Barnes, "Solving the pickup and delivery problem with time windows using reactive tabu search," *Transportation Research*, vol. 34, no. 2, pp. 107 – 121, 2000.

[5] J.-F. Cordeau, "A branch-and-cut algorithm for the dial-a-ride problem," *Oper. Res.*, vol. 54, no. 3, pp. 573–586, 2006. [Online]. Available: http://dx.doi.org/10.1287/opre.1060.0283

[6] R. Baldacci, E. Bartolini, and A. Mingozzi, "An exact algorithm for the pickup and delivery problem with time windows," *Operations Research*, vol. 59, no. 2, pp. 414–426, 2011. [Online]. Available: http://dx.doi.org/10.1287/opre.1100.0881

[7] A. Bettinelli, A. Ceselli, and G. Righini, "A branch-and-price algorithm for the multi-depot heterogeneous-fleet pickup and delivery problem with soft time windows," *Mathematical Programming Computation*, vol. 6, no. 2, pp. 171–197, 2014. [Online]. Available: http://dx.doi.org/10.1007/s12532-014-0064-0

[8] B. Bernay, S. Deleplanque, and A. Quilliot, "Routing on dynamic networks: GRASP versus genetic," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014.*, 2014, pp. 487–492. [Online]. Available: http://dx.doi.org/10.15439/2014F52

[9] J.-F. Cordeau, G. Laporte, and S. Ropke, *The Vehicle Routing Problem: Latest Advances and New Challenges*. Boston, MA: Springer, 2008, ch. Recent Models and Algorithms for One-to-One Pickup and Delivery Problems, pp. 327–357. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-77778-8_15

[10] R. Baldacci, A. Mingozzi, and R. Roberti, "Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints," *European Journal of Operational Research*, vol. 218, no. 1, pp. 1 – 6, 2012.

[11] H. Akeb, A. Bouchakhchoukha, and M. Hifi, "A beam search based algorithm for the capacitated vehicle routing problem with time windows," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, Kraków, Poland, September 8-11, 2013.*, 2013, pp. 329–336. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6644021

[12] ——, *Recent Advances in Computational Optimization: Results of the Workshop on Computational Optimization WCO 2013, FedCSIS 2013*. Cham: Springer International Publishing, 2015, ch. A Three-Stage Heuristic for the Capacitated Vehicle Routing Problem with Time Windows, pp. 1–19. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-12631-9_1

[13] Q. Lu and M. M. Dessouky, "A new insertion-based construction heuristic for solving the pickup and delivery problem with time windows," *European Journal of Operational Research*, vol. 175, no. 2, pp. 672 – 687, 2006.

[14] C. Zhou, Y. Tan, L. Liao, and Y. Liu, "Solving the multi-vehicle pick-up and delivery problem with time widows by new construction heuristic," in *Proc. IEEE CISDA*, vol. 2, 2006, pp. 1035–1042. [Online]. Available: http://dx.doi.org/10.1109/ISDA.2006.253754

[15] H. Li and A. Lim, "A metaheuristic for the pickup and delivery problem with time windows," in *Proc. IEEE ICTAI*, 2001, pp. 160–167. [Online]. Available: http://dx.doi.org/10.1109/ICTAI.2001.974461

[16] S. N. Parragh, K. F. Doerner, and R. F. Hartl, "A survey on pickup and delivery problems," *Journal fur Betriebswirtschaft*, vol. 58, no. 1, pp. 21–51, 2008. [Online]. Available: http://dx.doi.org/10.1007/s11301-008-0033-7

[17] S. Ropke and D. Pisinger, "An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows," *Transportation Science*, vol. 40, no. 4, pp. 455–472, 2006. [Online]. Available: http://dx.doi.org/10.1287/trsc.1050.0135

[18] G. Pankratz, "A grouping genetic algorithm for the pickup and delivery problem with time windows," *OR Spectrum*, vol. 27, no. 1, pp. 21–41, 2005. [Online]. Available: http://dx.doi.org/10.1007/s00291-004-0173-7

[19] Y. Nagata and S. Kobayashi, *Proc. PPSN XI*. Heidelberg: Springer, 2010, ch. A Memetic Algorithm for the Pickup and Delivery Problem with Time Windows Using Selective Route Exchange Crossover, pp. 536–545. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15844-5_54

[20] M. Cherkesly, G. Desaulniers, and G. Laporte, "A population-based metaheuristic for the pickup and delivery problem with time windows and LIFO loading," *Computers & Operations Research*, vol. 62, pp. 23 – 35, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305054815000829

[21] P. Kalina and J. Vokrínek, "Parallel solver for vehicle routing and pickup and delivery problems with time windows based on agent negotiation," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2012, pp. 1558–1563. [Online]. Available: http://dx.doi.org/10.1109/ICSMC.2012.6377958

[22] J. Nalepa and M. Blocho, *Intelligent Information and Database Systems: Proc. 8th Asian Conference, ACIIDS 2016*. Heidelberg: Springer, 2016, ch. Enhanced Guided Ejection Search for the Pickup and Delivery Problem with Time Windows, pp. 388–398. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-49381-6_37

[23] Y. Nagata and S. Kobayashi, "Guided ejection search for the pickup and delivery problem with time windows," in *Proc. EvoCOP*, ser. LNCS. Springer, 2010, vol. 6022, pp. 202–213. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12139-5_18

[24] T. G. Crainic and M. Toulouse, "Parallel meta-heuristics," in *Handbook of Metaheuristics*, ser. International Series in Operations Research & Management Science, M. Gendreau and J.-Y. Potvin, Eds. Springer US, 2010, vol. 146, pp. 497–541. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-1665-5_17

[25] J. Nalepa and M. Blocho, "Co-operation in the parallel memetic algorithm," *International Journal of Parallel Programming*, vol. 43, no. 5, pp. 812–839, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10766-014-0343-4

[26] T. G. Crainic and H. Nourredine, "Parallel meta-heuristics applications," in *Parallel Metaheuristics: A New Class of Algorithms*, M. Gendreau and J.-Y. Potvin, Eds. Wiley, 2005, pp. 447–494. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-1665-5_17

[27] G. Senarclens de Grancy and M. Reimann, "Evaluating two new heuristics for constructing customer clusters in a vrptw with multiple service workers," *Central European Journal of Operations Research*, vol. 23, no. 2, pp. 479–500, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10100-014-0373-4

[28] R. Banos, J. Ortega, C. Gil, F. de Toro, and M. G. Montoya, "Analysis of OpenMP and MPI implementations of meta-heuristics for vehicle routing problems," *Applied Soft Computing*, vol. 43, pp. 262 – 275, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494616300862

[29] Z. Manna, "Mathematical theory of partial correctness," *Journal of Computer and System Sciences*, vol. 5, no. 3, pp. 239 – 253, 1971. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022000071800351

[30] M. Blocho, "A parallel memetic algorithm for the vehicle routing problem with time windows," Ph.D. dissertation, Silesian University of Technology, 2013, (in Polish).

[31] S. Owicki and L. Lamport, "Proving liveness properties of concurrent programs," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 455–495, 1982. [Online]. Available: http://doi.acm.org/10.1145/357172.357178

[32] Z. Czech, *Introduction to Parallel Computing*. PWN, 2013.

[33] S. Wrona and M. Pawelczyk, "Controllability-oriented placement of actuators for active noise-vibration control of rectangular plates using a memetic algorithm," *Archives of Acoustics*, vol. 38, no. 4, pp. 529–536, 2013. [Online]. Available: http://dx.doi.org/10.2478/aoa-2013-0062

[34] J. Nalepa and M. Kawulok, "A memetic algorithm to select training data for support vector machines," in *Proc. of the 2014 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '14. New York, USA: ACM, 2014, pp. 573–580. [Online]. Available: http://doi.acm.org/10.1145/2576768.2598370

[35] K. Siminski, *Man–Machine Interactions 4: 4th International Conference on Man–Machine Interactions, ICMMI 2015 Kocierz Pass, Poland, October 6–9, 2015*. Cham: Springer International Publishing, 2016, ch. Memetic Neuro-Fuzzy System with Big-Bang-Big-Crunch Optimisation, pp. 583–592. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-23437-3_50

[36] J. Nalepa, M. Cwiek, and M. Kawulok, "Adaptive memetic algorithm for the job shop scheduling problem," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2015.7280409

[37] J. Nalepa and M. Blocho, "Adaptive memetic algorithm for minimizing distance in the vehicle routing problem with time windows," *Soft Computing*, vol. 20, no. 6, pp. 2309–2327, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00500-015-1642-4

[38] J. Nalepa and M. Kawulok, "Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs," *Neurocomputing*, vol. 185, pp. 113 – 132, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215019839

[39] M. Cwiek, J. Nalepa, and M. Dublanski, *Intelligent Information and Database Systems: Proc. 8th Asian Conference, ACIIDS 2016*. Heidelberg: Springer, 2016, ch. How to Generate Benchmarks for Rich Routing Problems?, pp. 399–409. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-49381-6_38

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 95–101
DOI: 10.15439/2016F411
ACSIS, Vol. 9. ISSN 2300-5963

An Economic Decision Support System based on Fuzzy Cognitive Maps with Evolutionary Learning Algorithm

Katarzyna Poczęta, Łukasz Kubuś, Alexander Yastrebov
Kielce University of Technology
al. Tysiąclecia Państwa Polskiego 7
25-314 Kielce, Poland
Email: {k.piotrowska,lkubus,jastri}@tu.kielce.pl

Elpiniki I. Papageorgiou
Technological Educational Institute (T.E.I.) of Central Greece
3rd Km Old National Road Lamia-Athens
Lamia 35100, Greece
Email: epapageorgiou@teiste.gr

***Abstract*—Fuzzy cognitive map (FCM) is a universal tool for modeling dynamic decision support systems. It can be constructed by the experts or learned based on data. FCM models learned from data are denser than those created by experts. We developed an evolutionary learning approach for fuzzy cognitive maps based on density and system performance indicators. It allows to select only the most significant connections between concepts and receive the structure more similar to the FCMs initialized by experts. This paper is devoted to the application of the developed approach to model an economic decision support system. The learning and testing process was accomplished with the use of historical data.**

I. Introduction

FUZZY cognitive map (FCM) is a directed graph for representing causal reasoning [9]. Nodes are variable concepts important for the analyzed problem and links are causal connections. FCM can be used for modeling decision support systems [12], [25], [26], e.g. for political decision making [4], artificial emotions forecasting [20] or prediction of work of complex systems [24].

Fuzzy cognitive map can be initialized based on expert knowledge. Experts select the concepts of the map and determine the weights of the connections between them (connection matrix). The second way to build the FCM model are learning algorithms [15]. Supervised [8], [18] and evolutionary learning algorithms [1], [3], [6], [12], [16], [22], [27] allow to determine the connection matrix based on historical data.

The resulting matrices are much denser than the models created by the humans. The density of the FCMs developed by experts is usually in the range of 30%-40% [23]. They choose only the most significant connections between concepts. FCMs with the smaller density are more readable for humans. Developing the learning algorithms that allow to build models in a manner more similar to human reasoning is an important part of research related to the fuzzy cognitive maps. A sparse real-coded genetic algorithm was proposed to utilize the density of the FCM model [23]. In [17], [19], a structure optimization genetic algorithm was introduces. A multi-objective evolutionary algorithm for learning maps with varying densities was analyzed in [7].

In [11], we propose a new evolutionary learning approach for fuzzy cognitive maps learning based on density and system performance indicators (SPI) analysis. It allows to obtain some compromise between data, density and significance of the connections. System performance indicators introduced by Borisov [5] and Silov [21] allow to analyze reliability of the FCM model and determine the total (direct and indirect) influence between concepts. The evaluation of the candidate FCMs is based on data error, density and the total influence between concepts. The obtained results based on the synthetic and real-life data generated from the reference matrices provided by experts proved that the developed approach allows to receive structure of the FCM model more similar to the reference object keeping the similar level of data error.

This paper is devoted to the use of fuzzy cognitive map with the developed evolutionary algorithm based on SPI in modeling economic decision support system. The aim of the analysis is to approximate the historical data and determine the influence between harmonized indexes of consumer prices for Poland [28]. The comparison of the developed approach with the standard one based on data error and the approach based on density was done. The learning process was accomplished with the use of Elite Genetic Algorithm (EGA) and Individually Directional Evolutionary Algorithm (IDEA) [10].

Section II presents fuzzy cognitive maps. Section III describes system performance indicators. The developed evolutionary algorithm for fuzzy cognitive maps learning is described in Section IV. In Section V, the results of experiments based on historical economic data are presented. Section VI contains the conclusion of the paper.

II. Fuzzy Cognitive Maps

Fuzzy cognitive map is a directed graph in the form [9]:

$$< X, W > \tag{1}$$

where $X = [X_1, ..., X_n]^T$ is the set of the concepts, W is the connection matrix describing weights of the connections, $w_{j,i}$ is the weight of the direct influence between the j-th concept and the i-th concept, taking on the values from the

range $[-1, 1]$. A positive weight of the connection $w_{j,i}$ means X_j causally increases X_i. A negative weight of the connection $w_{j,i}$ means X_j causally decreases X_i.

Fuzzy cognitive map can be used for modeling behavior of dynamic systems. The state of the FCM model is determined by the values of the concepts at the t-th iteration. The simulation of the FCM behavior requires an initial state vector. Next, the values of the concepts can be calculated according to the selected dynamic model. Simulations show the effect of the changes in the states of the map and can be used in a what-if analysis [2]. In the paper a popular dynamic model was used [22]:

$$X_i(t+1) = F\left(\sum_{j=1, j\neq i}^{n} w_{j,i} \cdot X_j(t)\right) \tag{2}$$

where $X_i(t)$ is the value of the i-th concept at the t-th iteration, $i = 1, 2, ..., n$, n is the number of concepts, t is discreet time, $t = 0, 1, 2, ..., T$. Transformation function $F(x)$ normalizes values of the concepts to a proper range. A logistic function is most often used [3], [22], [23]:

$$F(x) = \frac{1}{1 + e^{-cx}} \tag{3}$$

where c is a parameter, $c > 0$.

III. System performance indicators

System performance indicators allow to evaluate the structure of the FCM model e.g. by analysis the influence between concepts. Connection matrix W describes direct influence between concepts. The total influence $p_{j,i}$ between concepts means the maximum direct or indirect influence between concepts. To determine this system performance indicator the total causal effect path between concepts can be calculated [5], [21].

Algorithm for determining SPI contains the following steps [5], [21]:

1) First, connection matrix W with positive and negative direct relationships between concepts passes to matrix R size $2n \times 2n$ with positive relationships as follows:

$$\begin{array}{l} \text{if } w_{j,i} > 0 \\ \text{then } r_{2j-1,2i-1} = w_{j,i},\ r_{2j,2i} = w_{j,i} \\ \text{if } w_{j,i} < 0 \\ \text{then } r_{2j-1,2i} = -w_{j,i},\ r_{2j,2i-1} = -w_{j,i} \end{array} \tag{4}$$

2) Next, operation of transitive closure of the matrix R is used:

$$R^* = R \vee R^2 \vee R^3 \vee ... \tag{5}$$

where $\vee$ means maximum operation, R^k is calculated in accordance with the max-product composition:

$$R^k = R^{k-1} \circ R \tag{6}$$

3) Elements of the matrix R^* are transformed into matrix V as follows:

$$\begin{array}{l} v_{j,i} = max(r_{2j-1,2i-1}, r_{2j,2i}) \\ v'_{j,i} = -max(r_{2j-1,2i}, r_{2j-1,2i}) \end{array} \tag{7}$$

4) On the basis of the matrix V the total (direct and indirect) influence between the j-th concept and the i-th concept is calculated:

$$\begin{array}{c} \text{for } v_{j,i} \neq v'_{j,i} \\ p_{j,i} = sign(v_{j,i} + v'_{j,i}) max(|v_{j,i}|, |v'_{j,i}|) \end{array} \tag{8}$$

where $p_{j,i}$ takes value in the range of $[-1, 1]$.

5) The total influence between concepts can be used to determine other system performance indicators, for example the impact of the j-th concept on the system (9) and the impact of the system on the i-th concept (10):

$$\overrightarrow{P_j} = \frac{1}{n}\sum_{i=1}^{n} p_{j,i} \tag{9}$$

$$\overleftarrow{P_i} = \frac{1}{n}\sum_{j=1}^{n} p_{j,i} \tag{10}$$

IV. Evolutionary Learning Algorithm based on SPI

Evolutionary algorithms (like RCGA or EGA) can be applied to learn the FCM model (determine the weights of the connections between concepts) based on the available historical data. Each individual in the population is represented by a floating-point vector [22]:

$$W' = [w_{1,2}, ..., w_{1,n}, w_{2,1}, w_{2,3}, ..., w_{2,n}, ..., w_{n,n-1}]^T \tag{11}$$

where $w_{j,i}$ is the weight of the connection between the j-th and the i-th concept.

The individuals are decoded into the candidate FCMs and the response of every model are calculated based on the learning initial state vectors. The aim of the standard evolutionary algorithms for fuzzy cognitive maps learning is to minimize a total difference between the normalized historical data and the model response (data error), described as follows [3], [22]:

$$TE = \sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{i=1}^{n} |Z_i^p(t) - X_i^p(t)| \tag{12}$$

where $t = 0, 1, 2, ..., T$, T is the learning record length, $Z_i^p(t)$ is the reference value of the i-th concept at iteration t for the p-th learning record, $X_i^p(t)$ is the value of the i-th concept at iteration t of the candidate FCM started from the p-th initial state vector, $p = 1, 2, ..., P$, P is the number of the learning records.

Fuzzy cognitive maps learned with the use of methods based on data error properly perform the task of the input data approximation. However, the resulting connection matrices are denser than those initialized by experts [23]. Density of the FCM model can be expressed as a ratio of the number of non-zero weights and number of all possible non-zero weights according to the formula:

$$density = \frac{w_{non-zero}}{n^2 - n} \tag{13}$$

where $w_{non-zero}$ is the number of non-zero weights $w_{j,i}$, n is the number of the concepts.

To solve the problem of density the extensions of the standard evolutionary algorithms based on density analysis were proposed:

- sparse real-coded genetic algorithm [23],
- structure optimization genetic algorithm [17], [19],
- multi-objective evolutionary algorithm for learning FCM models with varying densities [7].

In [11], we introduced a new evolutionary approach for fuzzy cognitive maps learning that allow to determine the weights of the connections in the way similar to human reasoning. It is based on the analysis of data error, density and total influence between concepts in order to select only the most significant connections. Experiments performed with the use of synthetic and real-life data (generated from the reference FCMs) confirm that the resulting models are more similar to the reference systems.

The developed algorithm has the following objectives [11]:

- to minimize the data error (12),
- to minimize the FCM model density (13),
- to maximize the ratio of the number of significant total influences between concepts to the number of all possible influences described as follows:

$$PChRR = \frac{p_{relevant}}{n^2} \tag{14}$$

where $p_{relevant}$ is the number of total influences with the absolute value greater than 0.5 ($|p_{j,i}| > 0.5$).

To obtain some compromise between data error, density and significance of the influences between concepts, weighting method for determine the objective function was used [14]:

$$\begin{gathered} Error = a_1 \cdot TE + a_2 \cdot density \cdot TE + \\ + a_3 \cdot (1 - PChRR) \cdot TE \end{gathered} \tag{15}$$

where a_1, a_2, a_3 are parameters that meet the following condition:

$$\sum_{i=1}^{3} a_i = 1 \tag{16}$$

Density and total influence objectives are multiplied by the total error in order to lie in the same range.

Each candidate FCM is evaluated with the use of the following fitness function:

$$fitness(Error) = -Error \tag{17}$$

The developed approach consists of the following steps [11]:

STEP 1. Initialize random population.

Random initial population is generated and evaluated with the use of the fitness function. Each generated individual has density greater or equal to 20% and lower or equal to 50%.

STEP 2. Check stop condition.

If the number of iterations is greater than $iteration_{max}$ then stop the learning process.

STEP 3. Use evolutionary algorithm to generate new population.

In the simulation analysis of the proposed approach Elite Genetic Algorithm and Individually Directed Evolutionary Algorithm were used [10].

STEP 4. Analyze population.

The values from $[-0.05, 0.05]$ are rounded down to 0 as suggested in [22]. The total influence between concepts $p_{j,i}$ is calculated. The weight value $w_{j,i}$ is rounded down to 0 if the value of $p_{j,i}$ is in the interval $[-0.2, 0.2]$. Additionally, density is checked. Go to **STEP 2.**

STEP 5. Choose the best individual and test it.

A. Elite Genetic Algorithm

Elite Genetic Algorithm (EGA) [13] uses floating-point encoding as Real-Coded Genetic Algorithm (RCGA) and elite strategy. In the first step, random initial population is generated and evaluated. Next, temporary population T^t is created from current base population P^t by proportionate selection (roulette-wheel selection) with dynamic linear scaling of fitness function. Individuals of temporary population T^t are modified by Uniform Crossover operator and Non-Uniform Mutation operator [13]. The Uniform Crossover uses a fixed mixing ratio between two parents called *exchange probability* and usually it is equal to 0.5. Non-Uniform Mutation (NUM) operator keeps the population from stagnating in the early stages and decreases the range of mutation in later stages of evolution. In the last step, the temporary population T^t becomes new base population P^{t+1} after evaluation of the temporary population individuals.

B. Individually Directed Evolutionary Algorithm

Individually Directional Evolutionary Algorithm (IDEA) [10] uses floating-point encoding as EGA expanded by the additional mutation direction vector (DV). The DV is used by the mutation operation and correction of mutation direction process in post-selection stage. Random initial population is generated and evaluated as for EGA. Next, the roulette-wheel selection with dynamic linear scaling of fitness function is used to create temporary population T^t. Next, Directional Non-Uniform Mutation operator (DNUM) is used to create the second population $T^{'t}$ based on the temporary population T^t [10]. Individuals of temporary population $T^{'t}$ are evaluated and the next base population is created with the use of post-selection. The individual of temporary population T^t is compared to the corresponding individual from temporary population $T^{'t}$. Better individual is selected for the next base population. If the mutated individual is worse than the corresponding individual, the corresponding element of directional vector of the primary individual is corrected (correction of mutation direction).

V. Experiments

To analyze the performance of the developed evolutionary algorithm for fuzzy cognitive maps learning historical data were used [28]. The aim of the experiments is to build the economic decision support systems that allows to approximate the input data and analyze the influence between harmonized indexes of consumer prices for Poland. The comparative analysis of the developed approach with the standard one based on data error and the approach based on density was done.

A. Evaluation criteria

To evaluate the resulting FCM models, two criteria were calculated:

1) similarity between the input learning data and the data generated by the FCM candidate [22]:

$$initial_{error} = \frac{1}{P \cdot T \cdot n} \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{n} |Z_i^p(t) - X_i^p(t)| \tag{18}$$

where $t = 0, 1, 2, ..., T$, T is the learning record length, $Z_i^p(t)$ is the reference value of the i-th concept at iteration t for the p-th learning record, $X_i^p(t)$ is the value of the i-th concept at iteration t of the candidate FCM started from the p-th initial state vector, $p = 1, 2, ..., P$, P is the number of the learning records.

2) generalization capabilities of the candidate FCM (similarity between the input testing data and the data generated by the FCM candidate) [22]:

$$behavior_{error} = \frac{1}{P \cdot T \cdot n} \sum_{p=1}^{P} \sum_{t=1}^{T} \sum_{i=1}^{n} |Z_i^p(t) - X_i^p(t)| \tag{19}$$

where $t = 0, 1, 2, ..., T$, T is the testing record length, $Z_i^p(t)$ is the reference value of the i-th concept at iteration t for the p-th testing record, $X_i^p(t)$ is the value of the i-th concept at iteration t of the candidate FCM started from the p-th initial state vector, $p = 1, 2, ..., P$, P is the number of the testing records.

B. Dataset

The analyzed dataset contains monthly harmonized indexes of consumer prices for Poland collected in the period from January, 2011 to December, 2015 [28]:

- X_1 – date,
- X_2 – overall Index Excluding Alcohol and Tobacco for Poland,
- X_3 – all Items Excluding Administered Prices for Poland,
- X_4 – all Items Excluding Fully Administered Prices for Poland,
- X_5 – all Items Excluding Mainly Administered Prices for Poland
- X_6 – overall Index Excluding Energy for Poland,
- X_7 – overall Index Excluding Education, Health, and Social Protection for Poland,
- X_8 – overall Index Excluding Energy, Food, Alcohol, and Tobacco for Poland,
- X_9 – overall Index Excluding Energy and Seasonal Food for Poland,
- X_{10} – overall Index Excluding Liquid Fuels and Lubricants for Personal Transport Equipment for Poland
- X_{11} – overall Index Excluding Energy and Unprocessed Food for Poland,
- X_{12} – overall Index Excluding Housing, Water, Electricity, Gas, and Other Fuels for Poland,
- X_{13} – overall Index Excluding Seasonal Food for Poland,
- X_{14} – overall Index Excluding Tobacco for Poland.

Historical data were normalized according to the following equation:

$$f(x) = \frac{x - min}{max - min}\ , \tag{20}$$

where x is an input numeric value, min is the minimum of the dataset, max is the maximum of the dataset.

Data were divided into sequences of 10 consecutive records. 3 sequences were used as the learning records, and 3 sequences were used as the testing records.

C. Learning parameters

The following parameters were used for the EGA algorithm:

- selection method: roulette wheel selection with linear scaling
- recombination method: uniform crossover,
- crossover probability: 0.75,
- mutation method: non-uniform mutation,
- mutation probability: 0.02,
- population size: 10,
- number of elite individuals: 2,
- maximum number of iterations: 500,

The following parameters were used for the IDEA algorithm:

- selection method: roulette wheel selection with linear scaling
- mutation method: directed non-uniform mutation,
- mutation probability: $\frac{1}{n^2-n}$
- population size: 10,
- maximum number of iterations: 500,

The standard approach based on data error was realized for the objective function described by the equation (12). Table I shows the parameters of the objective function (15) for the approach based on density. Table II shows the parameters of the objective function (15) for the approach based on density and SPI. 10 experiments were performed for every set of the learning parameters and the average values (Avg) and standard deviations (Std) were calculated.

TABLE I
PARAMETERS OF THE OBJECTIVE FUNCTION FOR THE APPROACH BASED ON DENSITY ANALYSIS

L.p.	a_1	a_2	a_3
1	0.9	0.1	0
2	0.8	0.2	0
3	0.7	0.3	0
4	0.6	0.4	0
5	0.5	0.5	0

D. Results

Figure 1 shows an exemplary economic decision support system based on fuzzy cognitive map with the developed evolutionary algorithm. Table III summarizes the results of the experiments with historical data obtained for the standard approach (STD), the approach based on density (DEN) and the

TABLE II
PARAMETERS OF THE OBJECTIVE FUNCTION FOR THE PROPOSED APPROACH

L.p.	a_1	a_2	a_3
1	0.8	0.1	0.1
2	0.1	0.1	0.8
3	0.7	0	0.3
4	0.4	0.3	0.3
5	0.3	0.4	0.3
6	0.3	0.3	0.4
7	0.6	0	0.4
8	0.33	0.33	0.33

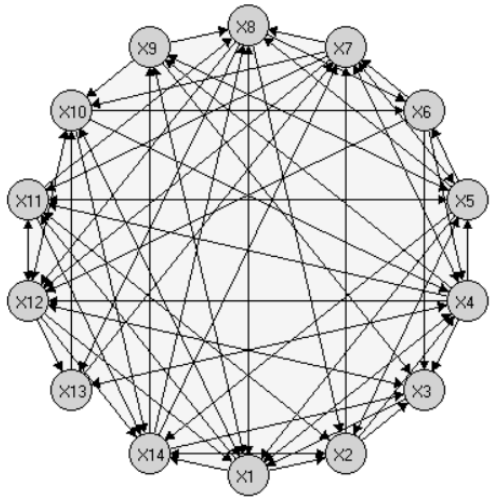


Fig. 1. Fuzzy cognitive map for economic decision support system

proposed approach (SPI). The obtained results show that the developed approach allows to approximate the economic data with satisfactory accuracy similar to the standard approach and the approach based on density. Figure 2 shows the sample results of testing of the resulting economic decision support systems.

The best solutions for the analyzed approaches are presented below. Table IV presents the resulting connection matrix obtained for the standard approach. Connection matrix for the approach based on density is shown in Table V. Table VI presents the connection matrix obtained for the proposed approach. The connection matrix for the standard approach has density 99%. The density for the second connection matrix (Table V) is equal 42%. The connection matrix for the evolutionary algorithm based on SPI has the lower density 40%. The FCM models obtained for the developed approach and method based on density are more readable and easier to interpret than the map obtained with the standard learning algorithm based only on data error.

For further analysis, impact of the j-th concept on the system (9) and impact of the system on the i-th concept (10) were calculated and presented in Table VII. The indicators obtained for the algorithm based on SPI have the highest absolute values. The developed approach allows to keep only

TABLE III
EXPERIMENTAL RESULTS WITH HISTORICAL DATA

Approach	Method	Parameters	$initial_{error}$ Avg ± Std	$behavior_{error}$ Avg ± Std
STD	IDEA		0.0427 ± 0.0001	0.0411 ± 0.0001
STD	EGA		0.0428 ± 0.0001	0.0411 ± 0.0001
DEN	IDEA	1	0.0427 ± 0.0001	0.0411 ± 0.0001
		2	0.0428 ± 0.0001	0.0411 ± 0.0001
		3	0.0428 ± 0.0001	0.0411 ± 0.0001
		4	0.0640 ± 0.0671	0.0628 ± 0.0686
		5	0.0428 ± 0.0001	0.0411 ± 0.0001
DEN	EGA	1	0.0433 ± 0.0002	0.0417 ± 0.0003
		2	0.0436 ± 0.0004	0.0419 ± 0.0005
		3	0.0435 ± 0.0003	0.0418 ± 0.0003
		4	0.0436 ± 0.0004	0.0420 ± 0.0004
		5	0.0435 ± 0.0002	0.0418 ± 0.0002
SPI	IDEA	1	0.0427 ± 0.0001	0.0411 ± 0.0001
		2	0.0445 ± 0.0053	0.0429 ± 0.0054
		3	0.0490 ± 0.0133	0.0475 ± 0.0137
		4	0.0444 ± 0.0050	0.0427 ± 0.0052
		5	0.0428 ± 0.0001	0.0411 ± 0.0001
		6	0.0469 ± 0.0071	0.0453 ± 0.0073
		7	0.0538 ± 0.0142	0.0524 ± 0.0145
		8	0.0428 ± 0.0001	0.0411 ± 0.0001
SPI	EGA	1	0.0568 ± 0.0227	0.0555 ± 0.0233
		2	0.0468 ± 0.0069	0.0452 ± 0.0070
		3	0.0596 ± 0.0237	0.0584 ± 0.0244
		4	0.0567 ± 0.0197	0.0553 ± 0.0203
		5	0.0516 ± 0.0120	0.0501 ± 0.0123
		6	0.0592 ± 0.0251	0.0579 ± 0.0257
		7	0.0482 ± 0.0078	0.0466 ± 0.0080
		8	0.0608 ± 0.0240	0.0595 ± 0.0246

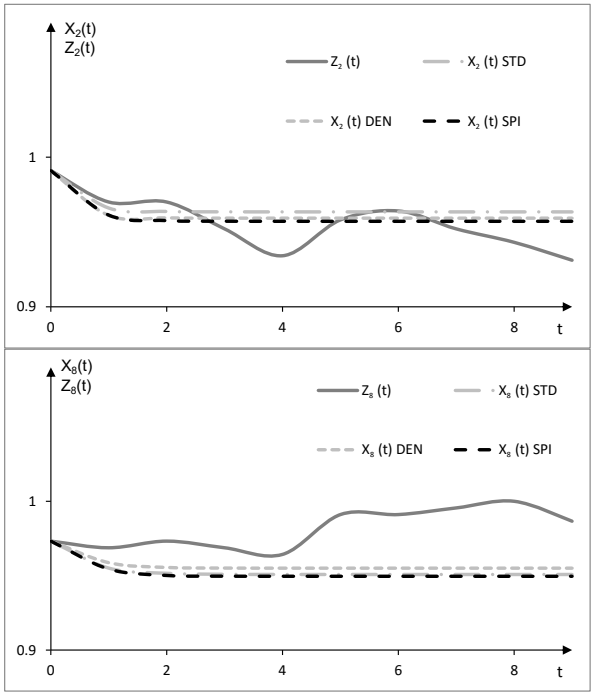


Fig. 2. Sample results of testing

the most significant connections between concepts and receive the structure for which all the concepts have a major impact

TABLE IV
CONNECTION MATRIX FOR THE FCM LEARNED WITH THE USE OF THE STANDARD APPROACH

$w_{j,i}$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
X_1	0	0.9	-0.15	0.98	-0.21	0.65	0.39	0.22	0.64	-0.17	0.56	0.42	0.82	0.21
X_2	0.97	0	0.34	0.18	0.92	0.28	-0.33	-0.4	-0.11	-0.74	0.69	-0.69	0.07	-0.35
X_3	0.64	-0.88	0	-0.43	0.44	0.83	0.47	0.07	-1	0.2	0.56	-0.13	0.36	0.66
X_4	-1	0.05	0.05	0	-0.89	-0.46	0.33	-0.14	0.99	0.91	0.04	0.08	-0.01	0.8
X_5	0.52	0.55	0.32	0.54	0	0.94	0.41	0.9	0.38	0.04	-0.67	0.86	0.71	-0.95
X_6	-0.66	0.77	0.21	0.62	-0.15	0	0.4	0.22	0.99	0.12	0.24	0.71	0.62	0.49
X_7	0.57	-0.73	-0.28	0.48	-0.61	-0.13	0	-0.24	-0.14	-0.14	-0.9	0.7	0.61	-0.24
X_8	-0.15	0.08	0.31	-0.1	0.76	0.88	-0.09	0	0.56	0.4	0	-0.2	0.57	0.73
X_9	0.42	0.03	0.8	-0.36	0.9	0.01	-1	0.85	0	0.94	-0.17	0.13	-0.66	0.13
X_{10}	-0.36	-0.4	-0.31	-0.62	0.05	-0.34	0.94	0.51	0.78	0	0.97	0.27	0.88	0.71
X_{11}	0.89	0.87	0.75	0.21	-0.13	-1	-0.17	-0.37	0.87	0.52	0	-0.18	-0.61	0.43
X_{12}	-0.05	-0.31	-0.41	-0.07	-0.34	0.97	0.64	-0.08	-0.46	0.94	0.89	0	0.03	0.52
X_{13}	-0.89	0.99	-0.39	0.42	0.93	0.53	0.96	-0.09	-0.25	0.08	-0.67	0.66	0	-0.55
X_{14}	-0.07	0.58	0.97	0.88	0.59	-0.57	-0.42	0.72	-0.52	-0.42	0.88	-0.34	-0.72	0

on the entire system and system has the significant impact on the concepts.

VI. CONCLUSION

The paper presents the application of fuzzy cognitive maps with the developed evolutionary algorithm for fuzzy cognitive maps learning to model economic decision support system. The presented approach is based on data error, density and system performance indicators analysis. The resulting FCM models are readable and easier to interpret. The proposed approach allows to keep only the most significant connections between concepts and receive fuzzy cognitive map for which all the harmonized indexes of consumer prices have a significant impact on the entire system and system has the significant impact on the indexes, keeping satisfactory accuracy of modeling of economic data.

REFERENCES

[1] G. Acampora, W. Pedrycz and A. Vitiello, "A Competent Memetic Algorithm for Learning Fuzzy Cognitive Maps," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 2397–2411, 2015, http://dx.doi.org/10.1109/TFUZZ.2015.2426311.

[2] J. Aguilar, "A Survey about Fuzzy Cognitive Maps Papers," *International Journal of Computational Cognition*,vol. 3 (2), pp. 27–33, 2005.

[3] S. Ahmadi, N. Forouzideh, S. Alizadeh, E. I. Papageorgiou, "Learning Fuzzy Cognitive Maps using Imperialist Competitive Algorithm," *Neural Computing and Applications*, vol. 26(6), pp. 1333–1354, 2015, http://dx.doi.org/10.1007/s00521-014-1797-4.

[4] A. S. Andreou, N. H. Mateou, and G. A. Zombanakis, "Soft Computing for Crisis Management and Political Decision Making: The Use of Genetically Evolved Fuzzy Cognitive Maps," *Soft Computing Journal*, Vol.9, Issue 3, pp. 194–210, 2005, http://dx.doi.org/10.1007/s00500-004-0344-0.

[5] V. V. Borisov, V. V. Kruglov, and A. C. Fedulov, *Fuzzy Models and Networks*, Publishing house Telekom, Moscow, 2004 (in Russian).

[6] A. Buruzs, M. F. Hatwagner, R. C. Pozna, and L. T. Koczy, "Advanced learning of fuzzy cognitive maps of waste management by bacterial algorithm," *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, pp. 890–895, 2013, http://dx.doi.org/10.1109/IFSA-NAFIPS.2013.6608518.

[7] Y. Chi and J. Liu, "Learning of Fuzzy Cognitive Maps with Varying Densities using Multi-objective Evolutionary Algorithms," *IEEE Transactions on Fuzzy Systems*, pp. 71-81, 2015, http://dx.doi.org/10.1109/TFUZZ.2015.2426314.

[8] A. Jastriebow and K. Poczęta, "Analysis of multi-step algorithms for cognitive maps learning," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 62, Issue 4, pp. 735–741, 2014, http://dx.doi.org/10.2478/bpasts-2014-0079.

[9] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no.1, pp. 65–75, 1986, http://dx.doi.org/10.1016/S0020-7373(86)80040-2.

[10] Ł. Kubuś, "Individually Directional Evolutionary Algorithm for Solving Global Optimization Problems - Comparative Study," *International Journal of Intelligent Systems and Applications (IJISA)*, Vol. 7, No. 9, 2015, str. 12-19.

[11] Ł. Kubuś, K. Poczęta, and A. Yastrebov, "A New Learning Approach for Fuzzy Cognitive Maps based on System Performance Indicators," *2016 IEEE International Conference on Fuzzy Systems*, Vancouver, Canada, pp. 1–7, 2016.

[12] N. H. Mateou and A. S. Andreou, "A Framework for Developing Intelligent Decision Support Systems Using Evolutionary Fuzzy Cognitive Maps," *Journal of Intelligent and Fuzzy Systems*, Vol. 19, Number 2, pp. 171–150, 2008.

[13] Z. Michalewicz, *Genetic algorithms + data structures = evolution programs*, Springer-Verlag, New York, 1996.

[14] K. M. Miettinem, *Nonlinear Multiobjective Optimization*, Kluwer Academic Publishers, Boston, 1999.

[15] E. I. Papageorgiou, "Learning Algorithms for Fuzzy Cognitive Maps - A Review Study." *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, vol. 42, no. 2, pp. 150–163, 2012, http://dx.doi.org/10.1109/TSMCC.2011.2138694.

[16] E. I. Papageorgiou, K. E. Parsopoulos, C. D. Stylios, P. P. Groumpos, and M. N. Vrahtis, "Fuzzy Cognitive Maps Learning Using Particle Swarm Optimization," *Journal of Intelligent Information Systems*, 25:1, pp. 95–121, 2005, http://dx.doi.org/10.1007/s10844-005-0864-9.

[17] E.I. Papageorgiou, K. Poczęta, and C. Laspidou, "Application of Fuzzy Cognitive Maps to water demand prediction," Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, Istanbul, pp. 1–8, 2015, http://dx.doi.org/10.1109/FUZZ-IEEE.2015.7337973.

[18] K. Poczęta and A. Yastrebov, "Analysis of Fuzzy Cognitive Maps with Multi-Step Learning Algorithms in Valuation of Owner-Occupied Homes," *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Beijing, China, pp.1029–1035, 2014, http://dx.doi.org/10.1109/FUZZ-IEEE.2014.6891587.

[19] K. Poczęta, A. Yastrebov, and E. I. Papageorgiou, "Learning Fuzzy Cognitive Maps using Structure Optimization Genetic Algorithm," *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Lodz, Poland, pp. 547–554, 2015, http://dx.doi.org/10.15439/2015F296.

[20] J. L. Salmeron, "Fuzzy cognitive maps for artificial emotions forecasting," *Applied Soft Computing*, vol. 12, pp. 3704–3710, 2012, http://dx.doi.org/10.1016/j.asoc.2012.01.015.

[21] V. B. Silov, *Strategic decision-making in a fuzzy environment*. Moscow: INPRO-RES, 1995 (in Russian).

[22] W. Stach, L. Kurgan, W. Pedrycz, and M. Reformat, "Genetic learning of fuzzy cognitive maps," *Fuzzy Sets and Systems*, vol. 153, no. 3, pp. 371–401, 2005, http://dx.doi.org/10.1016/j.fss.2005.01.009.

[23] W. Stach, W. Pedrycz, and L. A. Kurgan, "Learning of fuzzy cognitive maps using density estimate," *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, vol. 42(3), pp. 900–912, 2012, http://dx.doi.org/10.1109/TSMCB.2011.2182646.

TABLE V
CONNECTION MATRIX FOR THE FCM LEARNED WITH THE USE OF THE APPROACH BASED ON DENSITY

$w_{j,i}$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
X_1	0	0.54	0	0.47	0	0	0.28	0	0	0	0.98	0	0.99	0
X_2	-0.78	0	0	0	0	-0.21	0	0	0	0	0	0.32	0	0.89
X_3	0.56	0	0	0.77	0.9	0.53	-0.23	0.8	0	0.58	0.77	0.11	-0.32	0.36
X_4	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.97
X_5	0.5	0	0.91	0.79	0	0	0.56	0	0	0.81	0	-0.74	0.88	0
X_6	0	-0.49	0	0	0	0	0	0.91	0.29	0	0.69	0	0	0.79
X_7	0	1	0.9	0	0	0	0	0.07	0	0.33	0	0	0	0
X_8	0.77	0	0	0.34	0.46	0	0	0	0.47	0.55	0	0	0	0.59
X_9	0	0	0.86	0	0.91	0.99	0.66	0.97	0	0	0	0.71	0.62	0.92
X_{10}	0.35	0	0	0.88	0	-0.66	0.99	0	0	0	0	0	0	0
X_{11}	-0.9	0	0	-0.14	0	0.94	0.27	0	0.88	0	0	0.96	0.12	0
X_{12}	0	0.61	0	0	0	0.99	0	0	1	0.37	0	0	0	0
X_{13}	-0.27	0.87	-0.25	-0.87	0	0	0	0	0	0	0	0	0	0
X_{14}	0.6	0	0	0.45	0	-0.11	0	-0.57	0	0	0	0.87	0.42	0

TABLE VI
CONNECTION MATRIX FOR THE FCM LEARNED WITH THE USE OF THE PROPOSED APPROACH

$w_{j,i}$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
X_1	0	0.8	0	0.24	0	0	0.66	0.72	-0.89	0.56	0	0	-0.79	0
X_2	0	0	0	0.7	-0.22	-0.31	0	0	0	0	0.98	0	0	0
X_3	0	0	0	0	0.87	0	0	-0.34	0	0	0	0.2	0	0.99
X_4	-0.36	0.86	0	0	0	0.59	0	0	0	0.65	0.58	0	0	-0.31
X_5	0	0	0	0	0	0.99	-0.11	0	0	0	0	1	0	0
X_6	0	0	0.77	0	0	0	0.73	0	0	0.93	0	0	0.49	-0.83
X_7	0	0	0	0	-0.66	0	0	0	0.63	0	0	0	1	0.67
X_8	0	0	0	0	0	0	0	0	0.96	0	0	0.81	0	0
X_9	0.81	0	0	0	0.87	0	0	-0.11	0	0	0.71	0	0.87	-0.2
X_{10}	0	0	0	0	0.93	0	0	0	0	0	0	0.49	0.2	0.58
X_{11}	-0.33	0.79	0.48	0.77	0	0.65	0.36	0.38	0.52	0.62	0	0	0.98	0.89
X_{12}	0	0	1	0.98	0	0.9	0.23	-0.32	0.39	0	0	0	-0.31	0
X_{13}	0	0	0	0	0.47	0	0	0.94	0.91	0	0.08	-0.26	0	0.79
X_{14}	0.7	0	0	0	0	-0.35	0.72	0.97	0	0	0	0	0	0

TABLE VII
EXPERIMENTAL RESULTS WITH HISTORICAL DATA

Concept	STD		DEN		SPI	
	$\overrightarrow{P_j}$	$\overleftarrow{P_i}$	$\overrightarrow{P_j}$	$\overleftarrow{P_i}$	$\overrightarrow{P_j}$	$\overleftarrow{P_i}$
X_1	0.25	0.22	0.62	-0.22	-0.44	0.62
X_2	0.50	0.36	0.30	0.47	0.80	0.76
X_3	-0.37	-0.34	0.56	0.67	0.84	0.73
X_4	-0.14	0.36	-0.73	0.22	0.70	0.72
X_5	0.36	0.37	0.26	0.68	0.90	0.71
X_6	0.36	0.12	0.58	0.59	0.80	0.71
X_7	0.00	-0.26	0.45	0.58	0.78	0.65
X_8	0.43	0.35	0.59	0.71	0.79	0.78
X_9	0.22	0.26	0.82	0.58	0.81	0.75
X_{10}	0.13	0.38	0.49	0.56	0.85	0.67
X_{11}	-0.13	0.25	0.37	0.11	0.80	0.76
X_{12}	0.49	0.33	0.83	0.57	0.88	0.72
X_{13}	0.52	0.34	0.29	0.13	0.77	0.73
X_{14}	-0.23	-0.36	0.74	0.49	0.79	0.76

[24] G. Słoń, "Application of Models of Relational Fuzzy Cognitive Maps for Prediction of Work of Complex Systems," *Lecture Notes in Artificial Intelligence LNAI* 8467, Springer-Verlag, pp. 307–318, 2014, http://dx.doi.org/10.1007/978-3-319-07173-2_27.

[25] G. Słoń, "The Use of Fuzzy Numbers in the Process of Designing Relational Fuzzy Cognitive Maps", *Lecture Notes in Artificial Intelligence LNAI* 7894/Part 1, Springer-Verlag, pp. 376–387, 2013, http://dx.doi.org/10.1007/978-3-642-38658-9_34.

[26] G. Słoń and A. Yastrebov, "Optimization and Adaptation of Dynamic Models of Fuzzy Relational Cognitive Maps," in: S.O. Kuznetsov et al. (Eds.) *RSFDGrC 2011, Lecture Notes in Artificial Intelligence* 6743, Springer-Verlag, Heidelberg, pp. 95–102, 2011, http://dx.doi.org/10.1007/978-3-642-21881-1_17.

[27] E. Yesil and L. Urbas, "Big bang: big crunch learning method for fuzzy cognitive maps," *World Acad. Sci. Eng. Technol.*, vol. 71, pp. 815–8124, 2010.

[28] https://research.stlouisfed.org/fred2/downloaddata/ [20.04.2016]

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:

- AIPC'16—1st International Workshop on Advances in Image Processing and Colorization
- CANA'16—9th Computer Aspects of Numerical Algorithms
- CPORA'16—1st Workshop on Constraint Programming and Operation Research Applications
- IWCPS'16—3rd International Workshop on Cyber-Physical Systems
- MMAP'16—9th International Symposium on Multimedia Applications and Processing
- WSC'16—8th Workshop on Scalable Computing

1st Workshop on Constraint Programming and Operation Research Applications

THE aim of the CPORA-Workshop on Constraint Programming and Operation Research Applications is to bring together interested researchers from constraint programming/constraint logic programming (CP/CLP), operations research (OR) and artificial intelligence (AI) to present new techniques or new applications in decision support, combinatorial optimization, modeling and control processes arising in manufacturing, transportation, telecommunication, computer networks, logistic systems etc. and to provide an opportunity for researchers in one area to learn about techniques in the others. The aim of this workshop is share ideas, projects, researches results, models, experiences etc. associated with CP/CLP/OR/AI and to give researchers the opportunity to show how the integration of techniques from different fields can lead to interesting results on large and complex problems. Additionally, we would like to stimulate the communication between researchers working on different fields and practitioners who need reliable and efficient modelling and computational methods for industrial and business processes.

Contributions containing of both: the theoretical and practical results obtained in this area are welcome.

TOPICS

- Constraint programming/Constraint logic programming,
- Mathematical programming,
- Constraint Satisfaction Problem,
- Logic programming,
- Hybrid methods,
- Network programming,
- Petri-Nets,
- Knowledge methods,
- Soft computing (FL, GA, NN etc.),
- Answer Set Programming (ASP),
- The boolean satisfiability problem (SAT).
- Manufacturing,
- Multimodal processes management,
- Project management,
- Supply chain management,
- Modeling and planning production flow,
- Production scheduling,
- Multimodal social networks,
- Intelligent transport and passenger routing,
- Network knowledge modeling,
- Transportation networks.

EVENT CHAIRS

- **Bocewicz, Grzegorz,** Koszalin University of Technology, Poland
- **Sitek, Pawel,** Kielce University of Technology, Poland

PROGRAM COMMITTEE

- **Banaszak, Zbigniew,** Warsaw University of Technology, Poland
- **Bzdyra, Krzysztof,** Koszalin University of Technology
- **Gola, Arkadiusz,** Lublin University of Technology, Poland
- **Hajduk, Mikuláš,** Technical University of Kosice, Slovakia
- **Nielsen, Izabela Ewa,** Aalborg University, Denmark
- **Nielsen, Peter,** Aalborg University, Denmark
- **Relich, Marcin,** University of Zielona Gora, Poland
- **Terkaj, Walter**
- **Türkyılmaz, Ali,** Fatih University
- **Wikarek, Jarosław,** Kielce University of Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 107–113
DOI: 10.15439/2016F307
ACSIS, Vol. 9. ISSN 2300-5963

Simulation model of robotic manufacturing line

Grzegorz Gołda,
Silesian University of Technology
ul. Konarskiego 18A, 44-100
Gliwice, Poland
Email: grzegorz.golda@polsl.pl

Adrian Kampa,
Silesian University of Technology
ul. Konarskiego 18A, 44-100
Gliwice, Poland
Email: adrian.kampa@polsl.pl

Iwona Paprocka
Silesian University of Technology
ul. Konarskiego 18A, 44-100
Gliwice, Poland
Email: iwona.paprocka@polsl.pl

***Abstract*—The problem of production flow in the manufacturing line is analyzed. The machines can be operated by workers or by robots. Since breakdowns and human factors affect the destabilization of the production processes, robots are preferred to apply. The problem is how to determine the real difference in work efficiency between human and robot. Analysis of the production efficiency and reliability of the press shop lines operated by human operators or industrial robots are presented. This is a problem from the field of Operation Research area and Discrete Events Simulation (DES) method have been used. Two models have been developed including manufacturing line before and after robotization and taking into account stochastic parameters of availability and reliability of the machines, operators and robots. We apply OEE (Overall Equipment Effectiveness) indicator to present how the availability and reliability parameters influence over performance of the workstation, in particular in the short time and long time period. Also the stability of simulation model was analyzed.**

I. Introduction

THE industrial revolution caused the replacement of human labor by machinery. However, workers were still needed to handle and control the machines. Now we can see increasing use of automation and robotization, which replace human labor. Nowadays increased use of industrial robots can be observed especially for repetitive and high precision tasks (e.g. welding) or activities of monotonous and demanding physical exertion. Industrial robots have mobility similar to human arm, and can perform various complex actions like a human. In addition, they do not get tired and bored. It is estimated that thanks robotization, many companies obtained the reduction of the production cost by 50%, the increase of productivity by 30% and the increase of utilization by more than 85% [1].

However, the introduction of robotization requires incurring high costs, therefore robotization will be profitable only in certain circumstances, including, a high level of production, work with repetitive and precision tasks with ensuring the safety and health at work. Such conditions occur in the automotive industry and there the most robots are used.

The problem is how to determine a real difference in work efficiency between human and robot. The aim of the study is to develop a methodology, which allows to clearly define the throughput growth associated with the replacement of human labor by industrial robots. In order to assess the effectiveness of the robot application, we compare production uptime of humans and robots and calculate work efficiency with the use of the OEE indicator (Overall Equipment Effectiveness) and we use Discrete Event Simulation for verification.

II. Work efficiency and OEE Structure

Work efficiency and the use of the means of production can be expressed by using the OEE metric that depend on three factors: availability, performance and quality [2].

$$OEE=(Availability)x(Performance)x(Quality) \quad (1)$$

Availability is the ratio of the time spent on the realization of a task to the scheduled time. Availability is reduced by disruptions at work and machine failures.

$$Availability=\frac{avaliable\,time-failure\,time}{scheduled\,time} \quad (2)$$

Performance is the ratio of the time to complete a task under ideal conditions compared to the realization in real conditions or the ratio of the products obtained in reality to the number of possible products to obtain under ideal conditions. Performance is reduced (loss of working speed) by the occurrence of any disturbances e.g. human errors.

$$Performance=\frac{ideal\,cycle\,time}{real\,cycle\,time} \quad (3)$$

Quality is expressed by the ratio of the number of good products and the total number of products.

$$Quality=\frac{good\,products}{overall\,products} \quad (4)$$

The number of good quality products is a random variable, which can be described by a normal distribution

with standard deviation sigma. Quality levels are determined for ranges of the standard deviation sigma. In traditional production systems, level of 3 sigma is considered to be sufficient. However, in the modern automated and robotic systems the level of 5-6 sigma is possible to achieve [3].

A. Availability and failures

The term of availability contains planned work time and unplanned events e.g. the disturbances at work and random machine failures. Any unplanned event causes that machines are unavailable and work efficiency decreases. The reliability of objects such as machines or robots is defined as the probability that they will work correctly for a given time under defined conditions of work. The most popular method for estimating reliability parameters uses theory of probability to forecast a value of failure-free time and repair time parameters, under the condition that a trend based on historical value of the parameter is possible to notice. The examples of using normal, exponential, triangular distributions to describe both failure and repair times are described in [4]. In the article [5], it is assumed that parameters of distributions describing failure-free times, in general, change with time. Basing on information about the number of failures and failure-free times in a number of periods of the same duration in the past, some methods of estimation unknown parameters for scheduling purpose can be proposed [6].

In practice, for description of reliability, in most cases the parameter MTTF (mean time to failure) is used, which is the expected value of exponentially distributed random variable with failure rate λ [7].

$$MTTF = \int_0^\infty t\, f(t)\, dt = \int_0^\infty t\, \lambda\, e^{-\lambda x}\, dx = \frac{1}{\lambda} \quad (5)$$

In the case of repairable objects the parameters MTBF (mean time between failures), and the MTTR (mean time to repair) are used.

$$MTBF = MTTF + MTTR \quad (6)$$

For complex systems, consisting of n serially linked objects, the resultant failure rate λ_s of the system is the sum of the failure rates of each element λ_i:

$$\lambda_s = \sum_{i=1}^{n} \lambda_i \quad (7)$$

or the system $MTBF_s$ is the sum of inverse $MTBF_i$:

$$\frac{1}{MTBF_s} = \sum_{i=1}^{n} \frac{1}{MTBF_i} \quad (8)$$

For the example of robotic line, presented in figure 1, we can use formula 8 with different failure parameters for machines $MTBF_{mi}$ and for robots $MTBF_{ri}$:

$$\frac{1}{MTBF_s} = \sum_{i=1}^{n} \frac{1}{MTBF_{mi}} + \sum_{i=1}^{n+1} \frac{1}{MTBF_{ri}} \quad (9)$$

Machinery failures affect the availability of means of production and may cause severe disturbances in production processes. Average availability can be calculated with formula 10.

$$Availability = \frac{MTBF}{MTBF + MTTR} \quad (10)$$

Therefore, the longer the production line is, the higher the failure rate of the whole system. In industrial environment, the machine failures are mostly random and are difficult to predict; therefore, we have used computer simulation for further research [8].

III. Robotic Factor in Manufacturing

Manufacturing lines consist of different numbers of specialized machines and human operators or robots for materials handling. Usually, operator is required for loading and unloading the machine and for transferring the product from one machine to next production stage. Robots can make that work faster and more regular then human operators, but how fast a robot can work?

There are some methods for robot motion planning described in [9] and [10]. These methods are based on the MTM (Method Time Measurement) or on the traditional time study concept and can be used for comparing the relative abilities of robots and humans. Dedicated computer software for robot movement planning can be also used. The outcome of each technique is a set of time values that can be used to compare human and robot productivity.

In industry, there are many different types of machine tools, and presses are the most robotized. The schema of typical robotic press line is presented in the Figure 1.

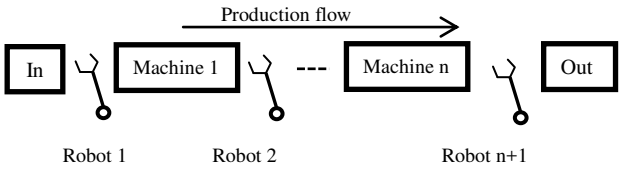


Fig. 1 The schema of robotic machine tending line

Robotized and automated lines are working very well but some problems with failures can occur. A failure of any elements of the line causes production stopping of the whole production line. Therefore, reliability of the components plays a key role for the productivity and utilization of manufacturing system. Consequently, in practice, the production lines consist mostly between 4 and 6 machines.

Modern industrial robots are characterized by a large precision of operation, high speed of motion and high reliability of work. These can be equipped with a various

tools and used to different works that are traditionally performed by human workers. It is important that the robots can work in conditions harmful to human health.

Some new-generation robots are equipped with various intelligent sensors, e.g. vision and pattern recognition systems, and they are able to adapt to changing conditions of external surroundings. New robots generation have also greater speed than older ones and can have important effect on robotic system performance.

Theoretically, robots can work 24 hours per day without any breaks, but human supervision of the production process and precise planning and scheduling of robot work are necessary for better performance [11]. Realized from time to time changes of tools and reprogramming require participation of an operator. Moreover, robot requires periodic maintenance service and inspection before each automatic run.

A. Robot reliability

For the first type of robots (Unimate) uptime was equal to MTBF=500 hours [10]. In article [12] the results of research on robots reliability at Toyota are presented. The reliability of first robot generation represents the typical bathtub curve. The next generation of robots was characterized by MTBF about to 8000 hours. Nowadays, robot manufacturers declare average MTBF=50000÷60000 hours or 20÷100 million cycles of work [13]. However, the robot's equipment is often custom made and therefore may turn out to be unreliable.

Some interesting conclusions from survey about industrial robots conducted in Canada [14] are as follows:

- Over 50 per cent of the companies keep records of the robot reliability and safety data,
- In robotic systems, major sources of failure are software failure, human error and circuit board troubles from the users' point of view,
- The most common range of the experienced MTBF is 500-1000 hours,
- Most of the companies need about 1-4 hours for the MTTR of their robots,

In the book [1] the approximate efficiency of robotic application versus manual application was compared. The efficiency of manual machine tending is about 40-60% and for robotic machine tending is about 90% (not including time for changeover setup equipment). However, detailed values are dependent on the specifics of the real workstation.

IV. EXAMPLE–PRESS LINE WITH WORKERS AND ROBOTS

In order to analyze the presented problem the mechanical press line from enterprise X, has been taken into account. Presses are often used in various production processes e.g. pressing, sheet metal forming etc. We have used Enterprise Dynamics software, which allows computer-modeling and simulation of discrete production processes with the use of human resources as well as robots.

In computer software used for production processes simulation the human factor is not sufficiently modelled. People are treated as quasi-technical elements of production system and they should operate in the same way as a machine. In practice, the human behaviour is unpredictable, thus it might help to explain why simulation models do not respond to the reality as it would be expected [15]. In the case of a manually-operated systems, a number of human factors (human errors) can lead to destabilization of the manufacturing process. Breaks for rest and higher requirements for Health and Safety at Work require a different way of working [16].

Computer models of lines operated by robots as well as by humans have been developed, taking into account the planned breaks at work and failure rates (Fig. 2, 4 and 5). The models contain the input (Source), storage buffers (Queue), machines, robots, human resources (Operators), output element for good quality products (Good parts) and one for poor quality products (Bad parts) and control elements (Availability Control, Schedule, MTBF, MTTR).

Unlimited supply of input materials and unlimited capacity for output products were assumed. Model contains some constrains, which are defined in objects parameters for example maximal buffer capacity.

The first model of robotic line without failures (Fig. 2), represents high production efficiency and achieve OEE above 90%, which is heavily dependent on the speed of robots movement.

Modern robots are characterized by increased speed and thanks to this, it is possible to obtain greater productivity. To examine the scale of this phenomenon several simulation with varying speed of the main robot axis, changing from 60 °/s to 210 °/s were conducted.

The relation between robot speed and robotic line productivity is presented in Figure 3. Initially, increased robot speed have allowed for a significant increase in production throughput. However, further increase of the robot speed does not increase the efficiency significantly.

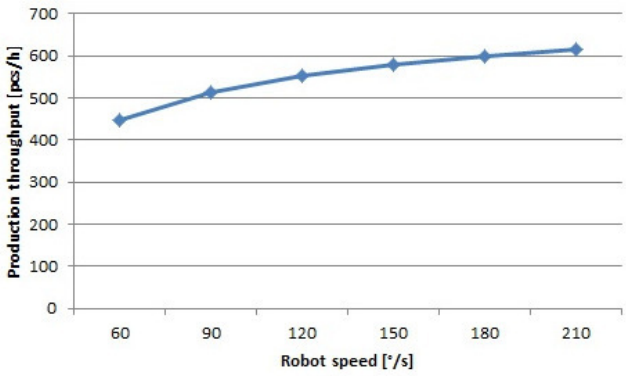


Fig. 3 Relation between productivity of robotic line and robot speed

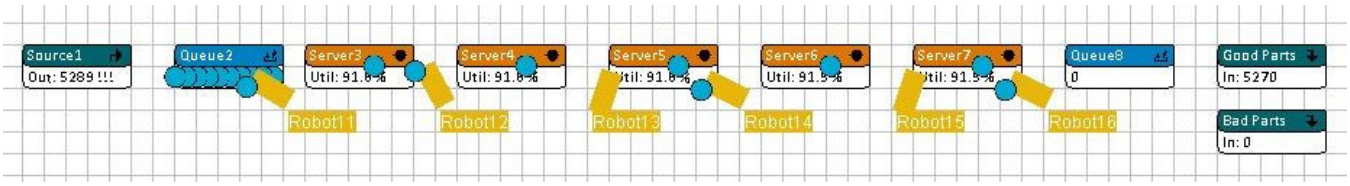


Fig. 2 Model of robotic line without failures after 8 hours of simulation

In practice reliability of machines and human errors are important issues. We take into account production parameters from enterprise "X". We assume that other employee can replace sick and absent worker, but it is impossible to replace broken machines and robots and they require repairing. In the case of failure occurrence the suspension of production on all machines in the line occurs. It has a huge impact on the performance of work, therefore we are taking into account failure parameters of machines and robots. The model include a number of work parameters: machine cycle time Tm=5 seconds; time of the line retooling 15 minutes (one time per shift) and reliability parameters for machines, MTBFm=500 hours and MTTRm=4 hours and robots, MTBFr=1000 hours and MTTRr=4 hours. The efficiency of the line and the speed of the robot 180 °/s equals to throughput rate about Pr=9.67 PCs/min, which is consistent with the data presented in [17]. Machine utilization equals about 80%.

In order to compare results we have also tested a model of manually-operated press line before robotization to determine differences in productivity. Manually operated line model is presented in Figure 4. The model consists of five machines, five operators and six buffer storages (Queue) in order to ensure continuity with the irregular performance of individual operators.

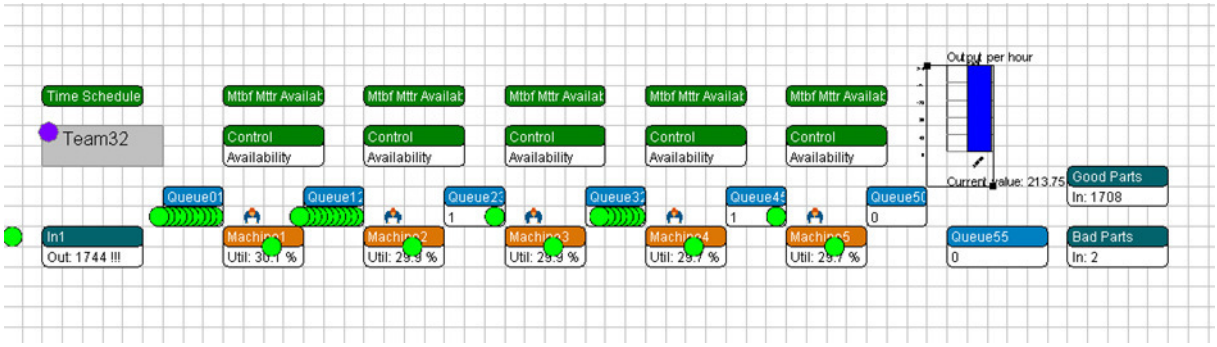


Fig. 4 Model of manually-operated line after 8 hours of simulation

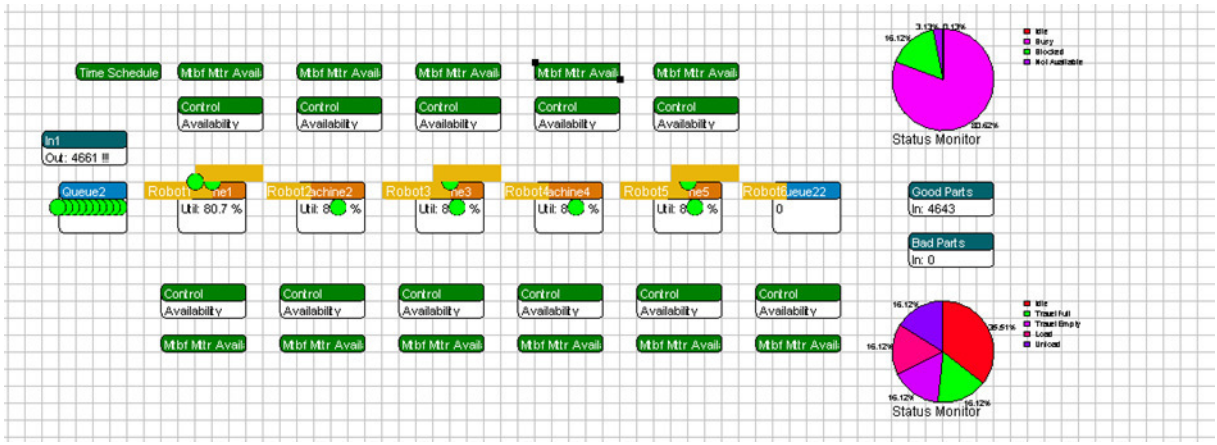


Fig. 5 Model of robotic line with failures after 8 hours of simulation

Parameters of operator were determined after time study and described by the normal distribution with average value of service time of 10 seconds and standard deviation of 2 seconds, which allows for implementation of nonuniformity in the work of the operators.

Assuming human unreliability on the basis of HEART (Human Error Assessment and Reduction Technique) for "routine and highly practiced rapid tasks involving relatively low level of skill", the nominal value of human error equals to 0.01 [18]. Therefore human errors rate can be described by parameters: MTBFh=8 hours and MTTRh=5 minutes. Taking into account the machine cycle parameters, Tm=5 seconds, the manually operated line should theoretically achieve production rate about P_h=4 PCs/min, but really the line achieved only about P_h=3.56 PCs/min. Utilization of machines equals around 30%.

The model of robotic line (Fig. 5) represents high production efficiency, which is heavily dependent on the speed of robots movement.

In addition, the stability of the production system and the impact of failure parameters on productivity and performance were analyzed in the similar manner as in the examples presented in [19]. A number of simulation experiments with different times of simulation runs were performed. Due to the random nature of a single failure process, a single simulation does not give complete picture of the situation. Therefore the experiment contains different number of simulations runs and simulation time from 8 hours to 6000 hours of work time. The trend lines of average production value for the manually operated line and for the robotic tended line are presented in Figures 6 and 7 respectively.

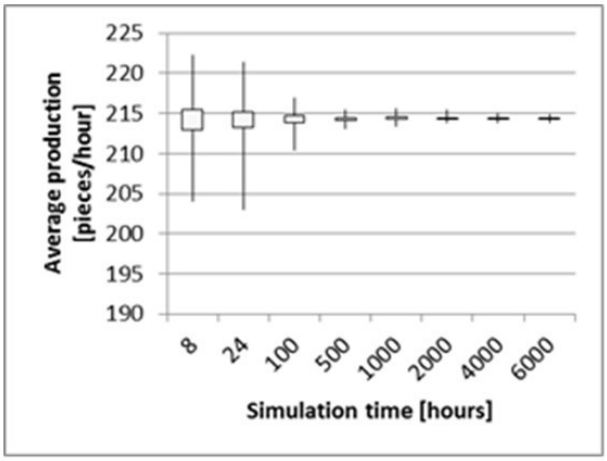


Fig. 6 The trend of average production value [pieces/hour] for manually-operated line (50 samples, confidence level 95%)

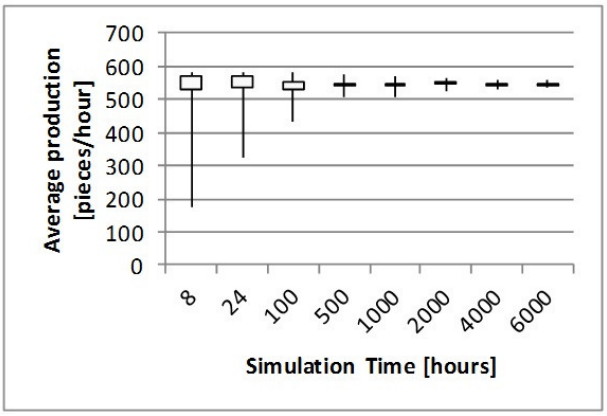


Fig. 7 The trend of average production [pieces/hour] value for robotic line (50 samples, confidence level 95%)

In the box and whisker plot, the average value of production is in the "box" range with confidence level of 95%. The "whiskers" show minimum and maximum range of production value.

The trends of average production value are more stable for longer simulation time. The model of robotic tended line show little difference with model of manually operated line. There are some outliers (the most extreme observations) represented by the minimum values of production that are connected with random failures and asymmetrical distribution with the left skewness can be observed (Fig. 8).

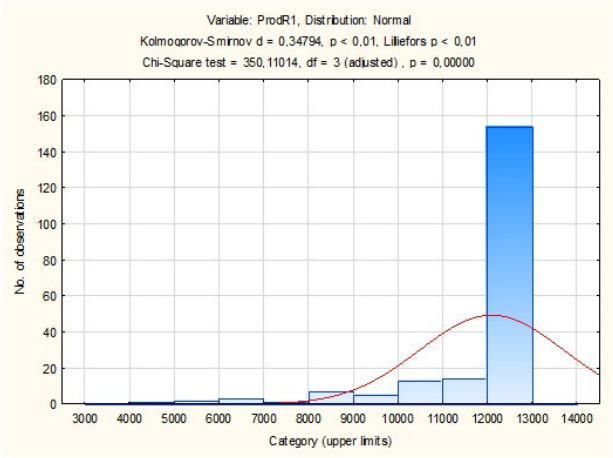


Fig.8 Histogram of production variable for robotic line after 24 hours of simulation

In other hands almost symmetrical normal distribution can be observed in the case of manually operated lines (Fig. 9).

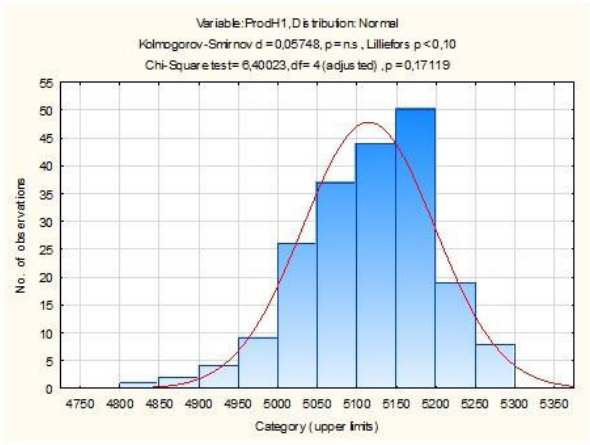


Fig. 9 Histogram of production variable for manual line after 24 hours of simulation

A number of fifty computer simulations for the simulation time from the range of 8 hour to 250 working days (one, two and three shifts and 250 working days per year) were run in order to observe the influence of long-term failures. For longer simulation time both models show decreased deviation and greater stability. Detailed results of the experiments are presented in the next section.

V. Simulation results

The production value P obtained from one simulation is a random variable that consists of several parameters. The random nature of the failures causes a significant dispersion of obtained values and relatively large standard deviation for confidence level α=0.95. The average production value P_{avg} of simulation experiments are summarized in table 1. Each experiment consists of fifty samples (simulation runs). The value *MaxLimit* determines the maximum possible production volume in a given period of time at the ideal working conditions for machine cycle time (Tm=5 seconds). Different reliability parameters in each column have been assumed in order to observe the influence of failures.

Since the model was build based on the OEE components, and contain parameters of availability, performance and quality, the production value from simulation can be directly used to calculate the OEE indicator.

$$OEE = \frac{Average\ production}{Maximal\ production\ limit} \quad (11)$$

The standard deviation shows the differences between the average value of production and the value of production achieved in each simulation run. For the robotic tended line, the values of standard deviation are also greater because of a much greater production volume and possibility of robots failures. This phenomenon can be explained that absent humans can be replaced but robots not.

TABLE I.
SIMULATION RESULTS FOR MANUALLY OPERATED AND ROBOTIC LINES (average production value P_{avg} in [PCs.] for 50 runs of simulation, α=0,95, MTBFm=500h, MTBFr=1000h, MTTR=4h and MTBFm=1000h MTBFr=2000h MTTR=4h)

	Human Operators	Robots	Human Operators	Robots
	MTBFm=500h MTBFr=1000h MTTR=4h		MTBFm=1000h MTBFr=2000h MTTR=4h	
	Time 8h		Time 8h	
Max Limit [PCs.]	5760		5760	
Average Production P_{avg} [PCs]	1681	4404	1713	4540
Standard deviation [PCs]	50.72	619.3	35.76	440.2
Relative deviation Δ	0.030	0.1406	0.0207	0.0969
OEE	**0,2918**	**0,7646**	**0,2974**	**0,7882**
	Time 24h		Time 24h	
Max Limit	17280		17280	
Average Production	5111	13300	5141	13521
Std. dev. [Pcs]	85.28	1432	84.09	967.6
Relative deviation Δ	0.0167	0.1077	0.0164	0.0716
OEE	**0,2958**	**0,7697**	**0,2975**	**0,7825**
	Time 2000h		Time 2000h	
Max Limit	1440000		1440000	
Average Production	427179	1094949	428765	1127235
Std. dev. [Pcs]	714.96	19331	662.4	12068
Relative deviation Δ	0.0016	0.0176	0.0015	0.0107
OEE	**0,2967**	**0,7604**	**0,2978**	**0,7828**
	Time 6000h		Time 6000h	
Max Limit	4320000		4320000	
Average Production	1281345	3279888	1286457	3379683
Std. Dev. [Pcs]	1207	27496	1306	24189
Relative deviation Δ	0.0009	0.0084	0.0010	0.0071
OEE	**0,2966**	**0,7592**	**0,2978**	**0,7823**

The relative deviation Δ indicates that the proportion of standard deviation to average production value is getting lower for long-time simulations. These effects are related to the occurrence of irregular failures in short-time simulations and to the almost regular occurrence of failures for long-time simulations. Thus simulation time should be greater than or equal to the largest value of the MTBF parameter.

Production throughput of robotic line has increased about 2.6 times comparing to the line before robotization.

The OEE related performance of a production line operated by a robot has improved by 48% comparing to a manually operated line. The OEE indicator equals to OOEh=29.66÷29,78% for humans and OEEr=75,92÷78,23% for robots, for 6000 hours of simulation, and correspond with the values assigned by the theory. Values calculated by theory are: availability of whole robotic system A=0,9085; performance P=0,8333; quality Q=0,9999. That gives OEE=75,7%. Reliability improvement can change the OEE score by about 2%. This shows that reliability parameters have significant influence on the productivity of the production system. Comparing the OEE factors for human operator and robot the greatest improvement is in the performance.

VI. CONCLUSIONS

The computer simulation of the simplified model of production line with machines, operators and robots with stochastic (short-time and long-time) reliability parameters allows for better representation and understanding of a real production process. The experiments confirm the advantage of application of robotic operated production lines comparing to manually operated lines. This is particularly to see in the case of work in three shifts for a long period of time. The work organization and robots synchronization play important role and therefore the efficiency of a production line operated by robots has improved OEE indicator by 46-48% comparing to a manually operated line.

Because of irregular work of human operators the buffers (queue) are needed for equalization of production flow and therefore loading (unloading) products from buffers results in low performance of human operators. Also breaks for rest results in lower OEE value.

However, in other cases of machine tools tending, the difference between human operator and robot is not so clearly to see even for long time simulations. The use of OEE factors allows comparing results from other manufacturing systems. The reality is that most manufacturing companies have OEE scores closer to 60%, but there are many companies with OEE scores lower than 45%, and small number of world-class companies that have OEE scores higher than 80%.

There are some place for improvement of availability, performance and quality. Availability depends on planned and unplanned breaks at work. Performance score depend on short machine cycle time and high robot speed. Quality depends on stability of manufacturing process parameters.

Obtained results can be used for detailed design of a robotic workcell and economic analysis, regarding labor costs and costs associated with the investments in robotization.

REFERENCES

[1] Glaser ,A., “Industrial robots”, Industrial Press, New York, 2009.

[2] Hansen, Robert C., “Overall Equipment Effectiveness”, Industrial Press, 2005.

[3] Barney, M., McCarty, T., “The new Six Sigma”, Prentice Hall Professional, New York, 2001.

[4] Gurel, S., Korpeglu, E. Akturk, M.S., “An anticipative scheduling approach with controllable processing times”. Computers and Operations Resesarch, Vol. 37, Issue 6, June 2010, pp. 1002-1013.

[5] Kempa W., Paprocka I., Kalinowski K., Grabowik C., “Estimation of reliability characteristics in a production scheduling model with failures and time-changing parameters described by gamma and exponential distributions”. Trans Tech Publications, (Advanced Materials Research; vol. 837 1662-8985), 2014, pp. 116-121.

[6] Paprocka I., Kempa W., Kalinowski K., Grabowik C., “A production scheduling model with maintenance”, Trans Tech Publications, Advanced Materials Research ; vol. 1036 1662-8985, 2014, pp. 885-890.

[7] Smith D.J. “Reliability, Maintainability and Risk. Practical methods for engineers”, Elsevier, Oxford, 2005.

[8] Paprocka I., Kempa W., Kalinowski K., Grabowik C.: “Estimation of overall equipment effectiveness using simulation programme”. Modern technologies in industrial engineering (ModTech2015), 17-20 June 2015, Mamaia, Romania. Materials Science and Engineering ; vol. 95 1757-8981. Institute of Physics Publishing, Bristol, 2015, s. 1-6, doi:10.1088/1757-899X/95/1/012155

[9] Genaidy, A.M., Gupta, T. ”Robot and human performance evaluation”. In “Human –Robot Interaction” Editor M. Rahimi, W. Karwowski, Taylor & Francis, London, 1992, pp. 4-15.

[10] Nof S. Y. “Handbook of industrial robotics”. John Wiley & Sons, New York, 1999.

[11] Kampa A., Planning and scheduling of work in robotic manufacturing systems with flexible production. Ed. B. Skołud: Hueristic Methods of Project and Production Scheduling. Journal of Machine Engineering, Vol. 12, No. 3, 2012, pp. 34-44.

[12] Sakai, H., Amasaka, K., “The robot reliability design and improvement method and the advanced Toyota production system", Industrial Robot: An International Journal, Vol. 34, 4, 2007, pp. 310–316

[13] Hägele, M. Nilsson, K., Pires J. N. ”Industrial Robotics”, in “Springer Handbook of Robotics”. Springer, Berlin, 2008

[14] Dhillon, B.S., Aleem, M.A.,"A report on robot reliability and safety in Canada: a survey of robot users ",Journal of Quality in Maintenance Engineering, Vol. 6 Issue 1, 2000, pp. 61–74.

[15] Baines, T., Mason, S., Siebers, P. O., Ladbrook, J., “Humans: the missing link in manufacturing simulation?”, Simulation Modelling Practice and Theory, No 12, 2004, pp. 515–526.

[16] Harriott, C. E., Adams, J. A., „Modeling Human Performance for Human–Robot Systems“, Reviews of Human Factors and Ergonomics, No 9, 94, 2013, http://rev.sagepub.com/content/9/1/94

[17] Kulanek, R. “Press Excelence Center COMAU, i.e. robotics press line at the world level”, available at www.robotyka.com 2014, (in polish, accessed 15.12.2015).

[18] Woods, D. D. “Modeling and predicting human error”, In J. Elkind, S. Card, J. Hochberg, and B. Huey (Eds.), Human performance models for computer-aided engineering Academic Press, 1990, pp. 248-274.

[19] Burduk, A., “Stability Analysis of the Production System Using Simulation Models”, Process Simulation and Optimization in Sustainable Logistics and Manufacturing, Springer, 2014, pp. 69-83.

3rd 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 as well as 2013 and 2015 IWCPS. The objective of the workshop is to assemble and develop a community with main interest in cyber-physical systems.

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

- **Babiceanu, Radu,** Embry Riddle Aeronautical University, United States
- **Ehrenberger, Wolfgang,** University of Applied Science Fulda, Germany
- **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
- **Saglietti, Francesca,** University of Erlangen-Nuremberg, Germany
- **Sanden, Bo,** Colorado Technical University, United States
- **Trybus, Leszek,** Rzeszow University of Technology, Poland
- **Vardanega, Tullio,** University of Padova - Dept. of Maths, 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. 117–124
DOI: 10.15439/2016F588
ACSIS, Vol. 9. ISSN 2300-5963

Comprehensive Observation and its Role in Self-Awareness; An Emotion Recognition System Example

Nima TaheriNejad, Axel Jantsch, David Pollreisz
Institute of Computer Technology, TU Wien
Gusshausstrasse 27-29, 1040 Vienna, Austria
{nima.taherinejad, axel.jantsch}@tuwien.ac.at

***Abstract*—Observation plays a crucial role in self-awareness. In many scenarios, such as the Observe-Decide-Act (ODA) loops, self-awareness is founded upon observations of the system. In other words, observation generates the understanding of the system from the status and behavior of its self and its environment. Although recently more focus has been put on comprehensive and competent observations, we believe that further attention and work is due, especially in the field of cyber-physical systems. Hence, in this paper, we discuss our position on various aspects of observation methods. In a short list, the major aspects are *Abstraction, Disambiguation, Desirability, Relevance, Data Reliability, Confidence, Attention*, and *History*. We elaborate and anticipate the potential of these factors in improving the quality of the observation of the system, decreasing the processing load of higher layers, increasing the reliability of decisions, and consequently the overall performance of the system. To put these aspects into perspective, we elaborate them in the context of their potentials in our emotion recognition system under development.**

I. Introduction

A COMMON requirement for many of the systems of today is an ability to perform correctly under a wide range of variation in their environment, as well as their internal states, parameters, applications and resources. Self-awareness is a feature that can enable these systems to show a robust and dependable behavior and meet their requirements.

Expecting 26 billion devices connected to the Internet of Things by 2020[1] with an exponential growth trend means that manual maintenance, fault diagnosis and repair for all will soon be impossible [1]. This necessitates an embedded awareness of the system regarding its own state so that it detects and mitigates occurring faults. Self-awareness has been applied to both hardware [1] and software [2]. Some of the applications which have been explored for the implementation of self-aware concepts (under this term or other terms such as adaptivity, autonomy, goal-oriented and so on), are mobile applications [3], cloud computing [4], networks [5], operating systems [6], web [7], multi-core resource managers [8] and adaptive and dynamic compilation environment [9], (cyber-physical) system-on-chip [1], and health monitoring [10].

[1]www.gartner.com/newsroom/id/2636073

Several definitions for Self-awareness can be found in the literature [11], [1]. For instance, in [11], it is defined as "a system's ability to obtain and maintain knowledge about its state, behavior and progress." In [12], [1], it is defined as the ability of the system to "monitor its behavior to update one or more of its components to achieve its goals". These definitions highlight the fundamental importance of monitoring, or observation, in self-awareness of the system.

Observation is more than collecting data. Higher levels of hierarchy, or the core in centralized units, are often burdened with a large load of monitored data and the respective processing [13]. On the other hand, local monitoring also may be delayed due to the global communication traffic [13]. Therefore, to be a useful basis of self- and environmental awareness it has to be a process that filters, analyses, selects, abstracts, assesses, and actively ignores or requests sensory data. Moreover, observation is not limited to sensory data, but rather a more general concept which includes other observations of the system from its own and its parts, regarding performance [8], functionality or other factors. In some works, such as [14], [1], this duty is encapsulated in a unit called "Inspection Engine" or "Introspective Sentient Unit" in [12].

In brief, the observation process transforms raw data into a high-quality description of the system about itself and its environment. In the following sections we describe the various aspects and activities involved in the observation process, especially that of a Cyber-Physical System (CPS), discuss examples in an emotion recognition setting and finally summarize the observation process, its importance and its potential contributions to a self-aware system.

The remainder of this paper is organized as the following: In the next section, a brief review of a self-awareness architecture is presented, followed by a short review of use-cases in Section III, where related works taking advantage of various aspects of a good observation are presented. Next, in Section IV, based on the aforementioned review of the existing literature, we present a summary of various aspects of a good observation scheme for self-aware CPSs. We clarify these concepts further in Section V, by producing an example in the context of an emotion recognition application which

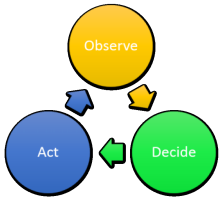


Fig. 1. Observe-Decide-Act (ODA) loop architecture for self-awareness.

we have under development. Finally, Section VI concludes the paper.

II. A Glance on the Architecture of Self-awareness

Many architectures have been proposed in order to achieve self-awareness. In CPSs with limited resources available for self-awareness, such as SEEC[2] [8] or Cyber-Physical System-on-Chip (CPSoC) [1], an Observe-Decide-Act (ODA) architecture is frequently used. ODA is one of the prominent architectures for the CPSs with a central supervisor or control unit. For example, in the SEEC framework, an application specifies the goal and the system, based on its observation of the performance, chooses the best action from a list of possible actions, in order to meet the goals. CPSoC is distinguished from a Multi-Processor System-on-Chip (MPSoC) due to its observation capabilities [12]. In CPSoC observation is done across various hardware and software layers to monitor performance, decide on a proper action and steer actuators.

As seen in Fig. 1, the ODA loop starts with observation, showing its importance. However, in a considerable number of CPSs, sensory data are directly or with minimal processing fed to and used by the decision-making unit. The processing is also often a simple interpretation, abstraction, or transformation of the raw data and scarcely entails any contextual information about the data, how it is obtained, how reliable it is and how it should be interpreted. This implies an inherent trust in the hard-coded interpretation schemes, health, precision and accuracy of sensors, and the obtained sensory data. Therefore, a fault or failure in the sensory system or the collected raw data can easily propagate to higher levels of decision-making and consequently into actions. The same stands for environmental or contextual changes which do not comply with the hard-coded interpretation schemes. Therefore, as a countermeasure learning and prediction is used in some self-aware systems [11], [1].

Self-awareness, however, is not necessarily achieved only through a centralized control or supervisory unit. In some works such as [15], [16], [11], it is a cooperative or emerging behavior of the group of agents or subsystems. Nevertheless, attention and context-aware observation play a very important role in creating self-awareness. Preden et al. [15] describe the concept of context-awareness under situation-awareness which consists of values and interpretations of a set of situation parameters. Situation parameters represent properties of a situation, processed and abstracted independently.

[2]SEEC is the name of the SElf-awarE Computing (SEEC) platform by Hoffman et al.[8]

III. Observation in the Literature

Traces of various aspects of a good observation can be tracked throughout the literature, although the terms, extent of usage, and thoroughness vary considerably. Thus, we base our position on a comprehensive observation of the literature, which provides us with good exemplars to follow in order to improve the performance of systems through improving their observation strategies. In this section, we briefly review some works in the literature which have emphasized or taken advantage of various aspects of a good observation scheme.

For example, in [8], the changes observed in the performance of the system are abstracted into heartbeats (an abstracted notion of time) which provide information about the execution of the application. This *abstraction* enables the framework to be adopted and used by various programs where the nature of goals are significantly different. All that is needed for this adoption is presenting the performance goal in the heartbeat abstraction notion. Moreover, available actions and reasoning in the decision-making process are abstracted into their effect on the speed of the system. Hence, SEEC serves as an example to show that abstraction does not necessarily need to be bottom up, and it can be used in top-down observations as well. We note that bottom-up abstraction often reflects the perception of the system from its environment, whereas top-down observations normally reflect the perception of the system about the performance of some parts of the system, or the system as a whole.

In [11], abstraction is done through online learning instead of predefined knowledge and rules. This increases flexibility and resilience of the overall system. In CPSoC [1], a virtual sensing platform is used both for the purpose of abstraction and *disambiguation* in the case of sensor fusions or faults and errors.

In [17] the authors have tried to create a unified *desirability* scale to compare and prioritize parameters of different nature which are directly related or comparable. A form of desirability scale is also found in [18].

Rinner et al. [11] briefly point to the selection of an object tracking algorithm which has “an acceptable level of robustness”, which approves of the importance of *confidence*, which is dependent on the *reliability* of other observations. Hoffman et al. [8], on the other hand, use the correlation of the recent past and near future, to assess the accuracy of the control scenario. Moreover, changes in this factor are considered in -and propagate to- the decision-making process. In CPSoC [1], a lifetime reliability characterization matrix is kept in the OS layer, which uses the data from reliability sensors in order to balance the workload of different units and thus increase the overall reliability during the lifetime.

The quality of sensors and measurements are in many systems unknown and cannot be guaranteed for the whole lifetime of the system. Therefore, *data reliability* analysis should always be an integral part of comprehensive observations.

In [11], for assessing the degree of the confidence of their observation (finding and tracking objects of interest), they use a measure of similarity. In this study, an existing database enables application of this procedure to obtain a figure of confidence. Alternatively, in some other systems *history* could replace this database or provide measures of confidence with other methods. However, self-assessment of the confidence by the system itself remains a major challenge in many cases. Specifically considering that a supervised assessment or validation of the performance of the system is not always possible, and often it is not desirable. Lack of history can increase the difficulty of autonomic acquisition of this aspect (confidence).

In [8], the *history* of recent observations are used to model the behavior of the system and predict whether a change in behavior is necessary, and if so, in what form. Typically, larger tracks of history can help in obtaining a better and more confident prediction of the future. However, maintaining a large memory is often challenging, in terms of storage and its respective costs, as well as the hardship and computation costs of processing a large amount of data. Therefore, it is important to use creative methods of storing data; that are compacting the data in such a way that necessary space decreases while valuable information is kept. Abstracting the data into the aspects mentioned above can help in achieving this goal.

IV. Aspects of a Comprehensive Observation

Building on the observation strategies in various related works (some of which were reviewed in Section III), in this section we present our position on various aspects of data interpretation which can lead to a good and comprehensive observation. Specific systems may not need or be able to afford all the required resources for such implementation as it will follow. Therefore, based on the details of the case, these aspects should be prioritized and used according to the case at hand.

A. Abstraction

1) Definition: Appropriate selection of the representation of the information in order to obtain compact knowledge, relevant to a particular purpose.

2) Description: Although collecting more data can help in improving the awareness and thus the performance of the system -since this relationship is not linear [14]- it does not guarantee that. To balance the processing load at various levels of the hierarchy, data need to be properly abstracted, which implies a meaningful mapping of the measured values to properties. A well-defined format is also crucial for a good abstraction, especially in distributed and hierarchical observation schemes [13].

3) Example: One of the parameters that is often measured in many e-health systems -and in our emotion recognition system as well- is the heartbeat rate. This value with all the potential noise and variations could be passed on to the central system for processing and decision making. Alternatively, these values can be mapped to some abstracted values (e.g., only five values of extremely low, low, normal, high, and extremely high) and thus transferred with smaller communication burden while decreasing the processing load of the decision-making unit.

4) Challenges: Predefined abstraction scenarios show little flexibility. While meaningful for a CPS in a well-defined environment with clear and stable objectives, it can cause an impediment for an autonomous system in following its evolving needs to adapt to new environment and situations.

5) Potential Solutions: As a solution, the system itself can create the most appropriate abstract properties based on unsupervised learning processes (for an example see [11]). This solution is significantly more challenging to implement and requires considerable resources. Nevertheless, it appears to be a prerequisite for truly autonomous and evolving systems.

B. Disambiguation

1) Definition: Remove uncertainty of meaning from measured data, or resolving conflicts arisen by different data.

2) Description: Despite a good abstraction scenario, the data can often be interpreted in more than one way or different parts of the system may have a different (conflicting) experience of the environment which could be natural or due to faults and failures. In either case, a functional, well-defined disambiguation strategy is a necessity. This prevents propagation of ambiguities or misinterpretations and thus facilitates decision making in the system.

3) Example: Heartbeat or temperature sensor may come from a chest belt, a smart watch or other sensors with various sensitivity and accuracy. Given information on data reliability (including accuracy or precision) may help in arbitrating non-matching values, e.g., heartbeat information from chest belt is more accurate than the one from a smart watch, hence the value of chest belt should be typically considered as the reference value, if the numbers are not matching.

4) Challenges: One of the challenges for disambiguation is again flexibility. For example, it is possible that the user is not wearing the chest belt and only noisy random values are observed, or for reasons of aging of the device or faults the accuracy of the chest belt has degraded. In such cases, the system should be able to change its arbitration and use alternative and more reliable values.

5) Potential Solutions: Parameters such as History, Confidence and Data Reliability of the measurements can help in identifying a need for a change in the disambiguation strategies and adapt the right alternative. A sudden drop in values can show taking off the chest belt (History), variations larger than usual, could imply some degradation (leading to decreased confidence) and loss or malfunction of some sensors in the chest belt constructing the final measurement could indicate less reliable measurements (data reliability).

On a higher level, and in a top-down flow of information, contextual awareness and predictions can provide meaningful inputs to the disambiguation unit as well. For example, is a low or high heartbeat accompanied by a low or high body temperature? Or, if the person has started or stopped a sport

activity, considerable and synchronized increase or decrease in heartbeat and temperature are expected (which do not reflect the feelings of the user).

C. Desirability

1) Definition: The quality of being inline with achieving one (or more) of the goals or expected outcomes of the system.

2) Description: Desirability is an evaluation of the system regarding the state of its own components and their alignment to its goals and expectations. When inferred states are marked as desirable or undesirable (possibly on a scale to distinguish between more and less desirable situations), a value system is implied that is rooted in the objectives and purpose of the system. If the system has only one sole purpose, the mapping of properties and states onto a desirability scale is less necessary. However, if the system pursues potentially contradicting goals a mapping onto a desirability scale is a useful intermediate step for resolving conflicts as well. A desirability scale serves as a unifying currency that allows for the comparison of otherwise unrelated properties and states.

3) Example: Let us assume the system is confronted with a discrepancy in the incoming data stream while on a low battery. The system has to decide if it is more important to preserve energy or to forward the detected anomaly for further processing. In such cases, a "1" or "0" bit on whether a discrepancy has happened or not, or the battery is full or low provides insufficient information for making a suitable and well-informed decision. If these events (observations) are mapped onto a desirability scale, this additional aspect of observation can show how undesirable a discrepancy has happened and how desirable would forwarding this problem be. Given the continuous level of charge in the battery (desirability of the available power), the system has further knowledge and flexibility in making an informed decision inline with its goals, for such situations where observations are not directly comparable or related.

4) Challenges: Creating an appropriate mapping of desirability for possible states of the components is sometimes very challenging. Particularly, when there is no direct link between the state of the component to be mapped to the desirability scale and the goals and expectations of the system.

5) Potential Solutions: One solution could be evaluating the effect of possible states of that component on achieving the goals of the system, via proxies and/or in isolation. However, this task itself (isolating the effect of the states of a non-directly linked component in achieving the goals), can be considerably challenging.

D. Relevance

1) Definition: The quality of being closely connected to/important for the matter at hand.

2) Description: Relevance has similarities with desirability, however, rather than states, it regards measurements and values or parameters and variables. Moreover, relevance regards smaller details in the system; that is, instead of considering alignment with the overall goals and expectations of the system, it considers the importance and connection of a measured value or parameter to a certain specific analysis in the system. This aspect can help in finding the right balance in the weight of a parameter in disambiguation, conflict resolution or decision-making as well as resources dedicated to that observation. This not only offloads the decision-making unit but also it enriches the observation unit by providing it with information on where more attention is necessary. Attention can affect how the raw data are obtained (such as the degree of redundancy, or frequency of data collection), or how it is processed, e.g., using precise methods requiring larger resources or less precise but lighter methods (such as approximate computing).

3) Example: Our preliminary measurements show that Anger and Sadness, for example, have a strong correlation with slow decrements of skin temperature, very weak correlation to the heartbeat and moderate to weak correlation to skin conductance, also known as Electro-Dermal Activity (EDA). This means that relevance of skin temperature to the two feelings is significantly higher than the other two. Hence, if the skin conductance and temperature are giving conflicting signals, skin temperature will be given greater weight in determining whether the feeling is one of the two or not.

4) Challenges: Traditionally, relevance is set by the designer, at design time. However, the relevance could change by the way of extracting it from a larger set of data. Once the designer does that, program upgrades are released. Nonetheless, this does not take into account personalization of this factor regarding the biometrics of the particular user.

5) Potential Solutions: The relevance of a parameter could be initialized, or predefined. Ideally, however, the system can learn the trends and change this value based on its experience and adapt it to its current situation and the biometrics of the specific user.

E. Data Reliability

1) Definition: the extent to which a measuring procedure yields the same results on repeated trials.

2) Description: As important as the data themselves, is the knowledge about its thoroughness, accuracy and precision, which allows the system to perform (here to observe/monitor) within the expected limits and stated conditions. Here accuracy describes the systematic bias of the data values compared to the real values of the measured quantity[3], and precision denotes the random errors in repeated measurements under the same conditions[4]. Although the concept of measurement device precision and propagation of inaccuracies into the system through calculations is well established, more attention by CPS engineers is due in system design[5].

3) Example: Awareness of the system about the robustness of the given values and their data reliability in certain situations can help the system to choose the best method for

[3]In other words, accuracy is a measure of statistical bias.

[4]In other words, precision is a measure of statistical variability.

[5]Although in approximate computing this matter is well considered, we believe that it should receive more attention in exact computing as well.

each situation (in the case of multi-mode technology) without needing to have an in-depth analysis of the situation and speculations on the accuracy of the given systems. That is, e.g., relying primarily on the heartbeat values coming from the chest-belt for precise information, before the smart watch and other means. Of course, if the data reliability of the chest-belt has been set to lower values (compared to the smart watch or other means), or if a problem, conflict or discrepancy has been already observed, the data with the highest reliability should be considered as the main/primary reference for the values.

4) Challenges: Propagation of measurements errors through the system, specifically, when several post-processing steps are taken, can impact the data reliability of the final assessment and hence the decision of the system concerning its situation. Whether the information is current or updated when it reaches the decision unit, or when a decision is reached, is another concern regarding the data reliability.

5) Potential Solutions: One solution is taking parameters partaking in the data reliability and robustness of measurements into account during the system design. These parameters could be the type of sensors and measurements (and its inherent accuracy and precision limits), bias set-up and peripherals of the sensory system, type and size of post-processing, validity period of the measured data, and the health (functionality) of the sensors themselves which could be affected by aging, or the environment.

We also note that one method to increase the data reliability is redundancy. Redundancy allows to collect the same data in different ways, for instance with various sensors, and to compare the consistency of the different, independent sources. One method of checking consistency is following the trend of changes in a single set of data and comparing the values with possible or impossible variations in the trend and data values. The other method is comparing different data sources and analyzing them for their mutual consistency.

F. Confidence

1) Definition: the extent to which a procedure may yield the same results on repeated trials.

2) Description: Confidence has significant similarities to data reliability, however, rather than measured data (or propagated results based on closed-form analysis), it regards experiments and procedures where a comprehensive external verification of accuracy and precision is not performed. This determines the firmness or flexibility of the system concerning its learning, analysis and decisions. In other words, to what extent the learned trends, analysis and decisions should be relied upon, or alternative analysis should be prepared or further data processed.

3) Example: Whether the temperature sensor is attached to the body or not, is not a measure of reliability. The confidence of the system in its attachment to the body, however, can affect how the system reacts to values out of the norm. To this end, a top-down variant of consistency analysis can be used. That is, a comparison of the measured values with expectations. Body temperature hardly undergoes a large change within seconds (or minutes). Therefore, a sudden drastic change could show a consistency problem and decrease the confidence of the system with regard to the measured value[6]. However, how reliable is consistency checking for assessing the attachment of the sensor is itself subject to the question of reliability (which constitutes the confidence of the system in this procedure). Should the system rely on this analysis or use other methods, such as asking the user?

4) Challenges: Creating a confidence measure, assessing or improving it can be a very challenging task, especially when it is to be unsupervised or with minimum external inputs. To be more specific, creating a fairly reliable (unsupervised) self-assessment procedure can be extremely hard.

5) Potential Solutions: Confidence can be built up based on the reliability of the inputs and the history of analysis. That is, analyzing the consistency of decisions, previous analysis, and the outputs given the data reliability of the inputs. One method of checking confidence is following the trend of outputs obtained by different procedures given a single set of data and comparing the values with each other and if possible, with the correct or desired values. The confidence values, in turn, can drive, how heavy a given procedure weighs in the interpretation of the current situation.

G. Attention

1) Definition: Selective allocation of limited resources to specific tasks.

2) Description: It is neither meaningful nor efficient to always collect and process all possible data. Attention decides on which observations to allocate the available communication, computation, memory and energy resources. The observation process should be guided by an understanding of what data are required for a given purpose in a given situation. At the same time, attention allows a low effort, continuous scanning of all sensory input to detect unexpected events and anomalies. If the attention process suspects the arrival of important new data, it focuses the attention, i.e. allocates the resources, on their collection and analysis. From these considerations, we infer that attention is driven both bottom-up, by the incoming data, and top-down, by prevailing expectations and goals.

3) Example: Using attention the system can, based on aspects such as relevance for example, optimize the resource allocation for data collection. As we present the details of our measurements in Section V, EDA has some correlation (weaker or stronger) with all targeted emotions. Therefore, instead of processing all measured data, EDA can be monitored and once an activity is observed there -based on the relevance to the suspected feeling- other processes can be activated selectively. Thus reducing the processing load of the system and optimizing resource utilization.

4) Challenges: In order to trigger attention in the observation process, the system should be able to detect certain trends and anomalies. A task which can be challenging given

[6]This, in turn, can trigger further processes which can determine whether this consistency problem is due to a normal environmental change (i.e., the sensor is detached) or a fault or failure.

the wide range of data that can be observed as well as trends or anomalies which can occur and need to be detected. However, using such methods for observatory data streams implies availability of very strict and limited resources which can render many prominent methods impractical. In addition, once an event is observed in a data stream which triggers further processing, other data which were not part of the focus of the system may be necessary and potentially missed.

5) Potential Solutions: A potential countermeasure for missing unfocused data when an event is detected and change of attention is triggered, could be using local buffers. Deciding the length of this buffer and its required resources can impose certain restrictions on the designer. Various learning methods and smart anomaly detection systems are potential solutions for detecting events in the data stream. However, we note that creating efficient methods with extremely limited resources is still a challenge.

H. History

1) Definition: Recording and studying a series of past events connected to an entity.

2) Description: History allows extraction of statistics from a series of observations; assessing the performance of the system over time, its improvements or deterioration, assessment of the quality of sensory data over time, and changes of the environment over time. Based on such historical data and analysis, the system is given the capability to predict its own performance in the near future, the expected failure of sensors and actuators, and future trends in the environment. In turn, this understanding can motivate the system to evolve its own goals, strategies, and tactics in order to adapt to these observations and predictions.

3) Example: History can play a crucial role in many different aspects, such as confidence (through consistency analysis, as mentioned above). It can increase relevance by establishing new correlations between factors that were not predetermined, however, their correlation is discovered during the process. In other words, what feelings correspond better to which measurements. For example, is happiness better associated with skin temperature or skin conductance? More importantly, history can help in determining the standard state of the system and identify potential abnormalities, which is a parameter tightly coupled with predictions. For example, history can establish what is the normal heartbeat rate of a person. This sets the expectation of the system on the heartbeat rate and its changes for each feeling.

4) Challenges: The greatest challenge of this aspect is the limited available memory. How to abstract, compact, and archive the data are some of other problems which often are affected by the limited memory as well.

5) Potential Solutions: By history, we do not necessarily imply an explicit bank of historical data. History can be inherent to some other parameters and implied rather than explicit with formal appearances. An instance is the average value or the trends of change and evolution of a parameter which may not include a bank of historical data to back the trend but does represent history. Another instance could be events and trends of their occurrence rather than the actual data. Periodically compacting and forgetting (eliminating old or less relevant data) are some other potential solutions.

V. Comprehensive Observation for Our Emotion Recognition System

To validate helpfulness of a comprehensive observation as discussed in Section IV, we plan to implement those aspects in an emotion recognition system. In this section, we try to clarify these concepts further by the way of examples in our design for our emotion recognition system under the development. Here, we suggest how the aforementioned aspects can be tailored, taken advantage, implemented and improve the "observe" unit of the system and consequently overall performance of the system.

A. Set-up

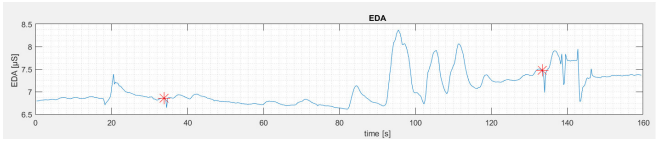
For our emotion recognition system, we use Empatica, "E4 Wristband" smart watch which has the following sensors embedded: A photoplethysmogram (PPG) to measure Blood Volume Pulse (BVP), Heartbeat and Heart Rate Variability (HRV), Electr-Dermal Activity (EDA) sensor for skin conductance, infrared thermopile for skin temperature, and 3-axis Accelerometer for capturing motions.

For measurements, subjects were asked to wear the smart watch and remain seated for one minute, so that we could obtain their baseline. To solicit Sadness, Happiness, Anger and Fear, short videos were played -in a random order- for the subjects. Afterwards, the subjects were asked to remain seated for one more minute (to get back to their baseline), while filling a self-assessment form, reporting the emotion they felt and how strongly they felt it.

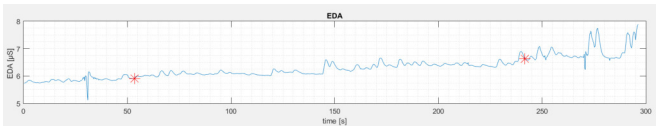
Measured data was collected and plotted in Matlab. In the future, this data needs to be post-processed and the emotions extracted on some hardware with very limited resources (such as a smartphone). For this reason, limited available resources is one of the primary constraints in this system, and we will describe how we plan to exploit comprehensive observation to increase the efficiency of the system.

B. Preliminary Measurements

Four participants, male, between the age of 20 and 25, took part in our preliminary measurements. Fig. 2 shows a sample, namely the measured EDA signal, during two different experiments: happiness and sadness. We observe that when the subject experiences the feeling during the video (between the star marks), his EDA changes with two different trends. For happiness, this trend appears as (relatively) large repetitive peaks, whereas as for sadness this seems to be a slow increment. Studying other measurements obtained from other subjects, the absolute value/level of EDA seems to carry no considerable significance in identifying these emotions. Therefore, an algorithm needs to be developed which can identify such trends in the EDA signal and notify the 'Decide' unit (see Fig. 1) of the system.



(a) Happiness



(b) Sadness

Fig. 2. Skin conductance (EDA) Signal for (a) Happiness, and (b) Sadness.

One of the simplest solutions to this problem is to define templates based on our measurements and compare the incoming data with the template. In this case, on top of identifying a matching/non-matching trend, the degree of similarity of the data with the template (how matching are the data compared to the template), can be used as the *confidence* of that unit in identifying the trends it is looking for. Hence, if a trend in the data is identified (or rejected), and thus causes a conflict in higher levels of the system, by looking into the *confidence* aspect, the potential source of conflict can be traced and the conflict can be resolved.

Table I summarizes the trends we have observed in our measurements. In this table, the background color of cells shows the *relevance* of that trend to the respective emotion. The lightness of the cell background shows how likely it is that with such emotion, such a trend in the measured parameter may be observed. To associate a level of *relevance* to each factor we have used the likelihood or unanimity of occurrence in our experiments. For example, in our measurements -unanimously- shortly after a person sees a scary video, their heartbeat increases. Therefore, we can confidently associate this observation with this feeling, and consider it highly *relevant*. On the other hand, a peak in EDA was not observed in every case, due to which we are less confident in our conclusion regarding the relevance of a single short peak in skin conductance, to the fear. Therefore, we consider it slightly less determinant for this feeling, and hence less *relevant*.

Our primary measurements on heartbeat values also show that only its changes from the baseline are relevant (rather than its absolute value/level). So the heartbeat can be *abstracted* to four values of increasing, decreasing, rise and fall, and constant (as shown in Table I). The regular heartbeat rate however, varies with the person under study. This variation

TABLE I
SUMMARY OF OBSERVED TRENDS IN OUR MEASUREMENTS.

	Skin Conductance	Skin Temperature	Heartbeat Rate
Happiness	Repetitive Peaks	Slow Increment	Constant
Sadness	Slow Increment	Slow Decrease	Rise and Fall
Fear	A short Peak	Constant	Small Increment
Anger	No Correlation	Slow Decrease	Constant

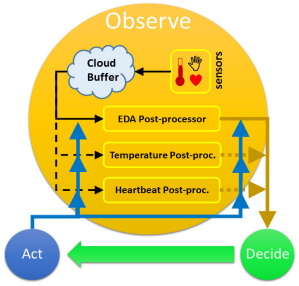


Fig. 3. Details of the "Observe" unit for our emotion recognition system.

should be considered when the person's emotion is being analyzed. To this end also, a small footprint learning method -based on the *history*- can be employed in the system. For the learning process, often feedback from higher levels (after categorization of emotions) are required and beneficial. Therefore, exchanging abstracted data needs to happen not only in a bottom-up manner but also in a top-down approach.

Based on such arguments and similar design analysis acquired during our preliminary measurements (a summary of which we have inserted in Table I), we propose the following system architecture for the "Observe" unit of our emotion recognition system. In this system we try to -whenever possible and beneficial- take advantage of as many observation aspects as possible.

C. System Architecture

For our system, we consider an ODA loop architecture. In this section, we further explain the details of the "Observe" unit in the architecture. As seen in Fig. 3, observation starts with collecting sensory data which are buffered on the cloud. This data need to be post-processed, in the following units.

To be efficient in communication and processing power, we take advantage of *attention*. That is, instead of continuously processing all data, they are periodically processed. To decrease the odds of missing an event, one data stream is considered as the default and is continuously processed. As mentioned in Section IV and seen in Table I, skin conductance (EDA) shows the most sensitivity to emotional experiences. Therefore, this stream is the default stream which is continuously processed in order to find potential events. Once an event is observed by the EDA post-processing unit, it triggers the "Decide" unit, which, based on the event and suspected/possible related emotion(s), can decide what other information is necessary. Accordingly, the action unit activates the respective post-processing unit(s) to acquire the relevant data from the cloud and process them. We note that the post-processing units in Fig. 3 are logical units and thanks to the *attention* based operation, the hardware itself could be shared.

Further details of the generic post-processor design are shown in Fig. 4. The most challenging unit in the post-processor is the matching unit where certain trends (according to Table I) are looked for and if found, *abstracted* data (abstracted following the trends categories listed in Table I) informs higher levels (decision unit) about them. Alongside

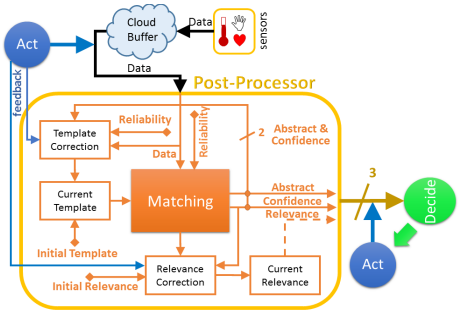


Fig. 4. Details of the post-processor unit.

the *abstracted* data, the *confidence* level of the findings can be sent to decision unit to facilitate the process of decision-making. *Confidence* information can be sent by the request of other units (such as when they need it for *disambiguation* or conflict resolution) or constantly with the *abstract*.

Relevance information on the other hand, undergoes very little and infrequent changes and therefore is sent only when a change in its value is observed in the post-processor unit (thus shown by a dashed line). Observing such a change necessitates keeping a *history* which is kept in the "relevance correction unit". This unit is initiated by the designer (represented as the background color of each cell in Table I) and updated by the feedback from higher levels of hierarchy (which includes information on the *confidence* of the given feedback) for each matched trend and taking into account the respective *confidence*s. A similar procedure transpires for updating the template which is used for finding the patterns. We note that for template correction -similar to matching unit- the *reliability* of sensory data is also used. A parameter which is provided by the designer at design time or provided automatically by the sensors, once attached.

VI. Conclusion

The importance of the observational process can hardly be overrated. In addition to the bare values, measured data have many contextual aspects that determine what it means and how it should be used. This host of information around the collected data is the basis of assessing the state of a system, its performance, its environment, and the influences of its actions.

Observation can be both a bottom-up and top-down process. The measured data are usually collected, processed, analyzed and abstracted bottom-up, while top-down expectations, hypotheses and needs greatly influence, steer, activate or block these bottom-up processes. Information redundancy and consistency analysis have to be used to assess confidence and reliability of data and lead to an appropriate assessment of the situation. Event-driven resource allocation (attention), triggered bottom-up or top-down, is another aspect which can considerably off-load the processing and communication.

After in-depth discussions on various aspects of a good and comprehensive observation, we described how we plan to use these concepts in our emotion recognition system under the development. Although these concepts can be employed, in "Decide" and "Act" units as well, since the role of a comprehensive observation is bolder in the "Observe" unit, we described the details of our comprehensive observation plan only for this unit of our system.

References

[1] N. Dutt, A. Jantsch, and S. Sarma, "Toward smart embedded systems: A self-aware system-on-chip (SoC) perspective," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 15, no. 2, p. 22, 2016.

[2] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[3] P. Mercati, A. Bartolini, F. Paterna, T. S. Rosing, and L. Benini, "A linux-governor based dynamic reliability manager for android mobile devices," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*. IEEE, 2014, pp. 1–4.

[4] B. Jennings and R. Stadler, "Resource management in clouds: Survey and research challenges," *Journal of Network and Systems Management*, pp. 1–53, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10922-014-9307-7

[5] P. eth ee Spathis and M. Bicudo, "ANA: Autonomic network architecture," *Autonomic Network Management Principles*, p. 49, 2011.

[6] L. Wanner, S. Elmalaki, L. Lai, P. Gupta, and M. Srivastava, "VarEMU: An emulation testbed for variability-aware software," in *Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2013 International Conference on*, 2013, pp. 1–10.

[7] J. Strassner, S.-S. Kim, and J. W.-K. Hong, "The design of an autonomic communication element to manage future internet services," in *Management Enabling the Future Internet for Changing Business and New Computing Services*. Springer, 2009, pp. 122–132.

[8] H. Hoffmann, M. Maggio, M. D. Santambrogio, A. Leva, and A. Agarwal, "SEEC: A framework for self-aware computing," MIT, Cambrige, Massachusetts, Tech. Rep. MIT-CSAIL-TR-2010-049, October 2010.

[9] W. Baek and T. M. Chilimbi, "Green: a framework for supporting energy-conscious programming using controlled approximation," in *ACM Sigplan Notices*, vol. 45, no. 6. ACM, 2010, pp. 198–209.

[10] J.-S. Preden, K. Tammemäe, A. Jantsch, M. Leier, A. Riid, and E. Calis, "The benefits of self-awareness and attention in fog and mist computing," *IEEE Computer, Special Issue on Self-Aware/Expressive Computing Systems*, pp. 37–45, July 2015.

[11] B. Rinner, L. Esterle, J. Simonjan, G. Nebehay, R. Pflugfelder, G. Fernandez Dominguez, and P. R. Lewis, "Self-aware and self-expressive camera networks," *Computer*, vol. 48, no. 7, pp. 21–28, 2015.

[12] S. Sarma, N. Dutt, P. Gupta, A. Nicolau, and N. Venkatasubramanian, "Cyberphysical-system-on-chip (CPSoC): A self-aware MPSoC paradigm with cross-layer virtual sensing and actuation," in *Proceedngs of the Design, Automation and Test in Europe Conference and Exhibition (DATE)*, Grenoble, France, March 2015.

[13] L. Guang, J. Plosila, J. Isoaho, and H. Tenhunen, "Hierarchical agent monitored parallel on-chip system: A novel design paradigm and its formal specification," *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*, vol. 1, no. 2, 2010.

[14] A. Jantsch and K. Tammemäe, "A framework of awareness for artificial subjects," in *Proceedings of the 2014 International Conference on Hardware/Software Codesign and System Synthesis*, ser. CODES '14. New York, NY, USA: ACM, 2014, pp. 20:1–20:3. [Online]. Available: http://jantsch.se/AxelJantsch/papers/2014/AxelJantsch-CODES.pdf

[15] J.-S. Preden, J. Llinas, G. Rogava, R. Pathma, and L. Motus, "On-line data validation in distributed data fusion," in *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IV: SPIE Defense, Security and Sensing*, T. Pham, M. A. Kolodny, and K. L. Priddy, Eds. SPIE - International Society for Optics and Photonics, 2013.

[16] J.-S. Preden, "Generating situation awareness in cyber-physical systems: Creation and excahnge of situational information," in *Proceedings of the 2014 International Conference on Hardware/Software Codesign and System Synthesis*. New York, NY, USA: ACM, October 2014.

[17] M. Sánchez-Escribano and R. Sanz, "Emotions and the engineering of adaptiveness," in *Procedia Computer Science: Conference on Systems Engineering Research*, vol. 28. Madrid, Spain: Elsevier, 2014, pp. 473–480.

[18] H. Hoffmann, "CoAdapt: Predictable behavior for accuracy-aware applications running on power-aware systems," in *Real-Time Systems (ECRTS), 2014 26th Euromicro Conference on*, July 2014, pp. 223–232.

8th Workshop on Scalable Computing

THE Workshop on Scale Computing (WSC) is a result of evolution in the world of computing. It originated (as Workshop on Large Scale Computing in Grids; LaSCoG) in 2005. Next, cloud computing became popular and, in response to this new trend, Workshop on Scalable Computing in Distributed Systems (SCoDiS) emerged. The two workshops (under a joint name LaSCoG-SCoDiS) have been organized till 2014 (information about past events can be found here). However, the world of large-scale computing continuously evolves. In particular, data-intensive computations (known as “Big Data") brough a completely new set of issues that have to be solved (in addition to those that exist since late 1990th and that still deserve our attention). Therefore we have decided to refresh the name of the event (to better represent the scope of interest). This is how the Workshop on Scalable Computing (WSC) came to being.

TOPICS

- General issues in scalable computing
  - Algorithms and programming models for large-scale applications, simulations and systems
  - Large-scale symbolic, numeric, data-intensive, graph, distributed computations
  - Architectures for large-scale computations (GPUs, accelerators, quantum systems, federated systems, etc.)
  - Data models for large-scale applications, simulations and systems
  - Large-scale distributed databases
  - Security issues for large-scale applications and systems
  - Load-balancing / intelligent resource management in large-scale applications, simulations and systems
  - Performance analysis, evaluation and prediction
  - Portals, workflows, services and collaborative research
  - Data visualization
  - On-demand computing
  - Virtualization supporting computations
  - Self-adaptive computational / storage systems
  - Volunteer computing
  - Scaling applications from small-scale to exa-scale (and back)
  - Computing for Big Data
  - Business applications
- Grid / Cloud computing
  - Cloud / Grid computing architectures, models, algorithms and applications
  - Cloud / Grid security, privacy, confidentiality and compliance
  - Mobile Cloud computing
  - High performance Cloud computing
  - Green Cloud computing
  - Performance, capacity management and monitoring of Cloud / Grid configuration
  - Cloud / Grid interoperability and portability
  - Cloud / Grid application scalability and availability
  - Economic, business and ROI models for Cloud / Grid computing
  - Big Data cloud services

EVENT CHAIRS

- **Ganzha, Maria,** University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland
- **Gusev, Marjan,** University Sts Cyril and Methodius, Macedonia
- **Paprzycki, Marcin,** Systems Research Institute Polish Academy of Sciences, Poland
- **Petcu, Dana,** West University of Timisoara, Romania
- **Ristov, Sashko,** University Sts Cyril and Methodius, Macedonia

PROGRAM COMMITTEE

- **Anderson, David,** University of California, Berkeley, United States
- **Bass, Len,** NICTA, Australia
- **Brodnik, Andrej,** University of Ljubljana, Faculty of Computer and Information Science, Slovenia
- **Camacho, David,** Universidad Autonoma de Madrid, Spain
- **D’Ambra, Pasqua,** ICAR-CNR, Italy
- **Filippone, Salvatore,** University Rome Tor Vergata, Italy
- **Gepner, Paweł,** Intel Corporation, United Kingdom
- **Gordon, Minor,** Software development consultant, United States
- **Goscinski, Andrzej,** Deakin University, Australia
- **Gravvanis, George,** Democritus University of Thrace, Greece
- **Grosu, Daniel,** Wayne State University, United States
- **Holmes, Violeta,** The University of Huddersfield, United Kingdom
- **Hsu, Ching-Hsien (Robert),** Chung Hua University, Taiwan
- **Kalinov, Alexey,** Cadence Design Systems, Russia
- **Karaivanova, Aneta,** IICT-BAS, Bulgaria
- **Kitowski, Jacek,** AGH University of Science and Technology, Department of Computer Science, Poland

- **Knepper, Richard,** Indiana University, United States
- **Kranzlmüller, Dieter,** Ludwig-Maximilians-Universität München (LMU), Germany
- **Kwiatkowski, Jan,** Wrocław University of Technology, Poland
- **Lang, Tran Van,** Vietnam Academy of Science and Technology, Vietnam
- **Lastovetsky, Alexey,** University College Dublin, Ireland
- **Legalov, Alexander,** Siberian Federal University, Russia
- **Luo, Mon-Yen,** National Kaohsiung University of Applied Sciences, Taiwan
- **Margaritis, Konstantinos G.,** University of Macedonia, Greece
- **Milentijevic, Ivan,** University of Nis, Serbia
- **Morrison, John,** University College Cork, Ireland
- **Nosovic, Novica,** Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
- **Olejnik, Richard,** CNRS - University of Lille I, France
- **Ouedraogo, Moussa,** Public Research Centre Henri Tudor, Luxembourg
- **Rak, Massimiliano,** Seconda Università di Napoli, Italy
- **Schikuta, Erich,** University of Vienna, Austria
- **Schreiner, Wolfgang,** Johannes Kepler University Linz, Austria
- **Shen, Hong,** University of Adelaide, Australia
- **Song, Ha Yoon,** Hongik University, South Korea
- **Stankovski, Vlado,** University of Ljubljana, Slovenia
- **Talia, Domenico,** University of Calabria, Italy
- **Telegin, Pavel,** JSCC RAS, Russia
- **Trystram, Denis,** Grenoble Technical University, France
- **Tudruj, Marek,** Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland
- **Tvrdik, Pavel,** Faculty of Information Technology, Czech Technical University in Prague, Czech Republic
- **Vazhenin, Alexander,** University of Aizu, Japan
- **Wei, Wei,** School of Computer science and engineering, Xi'an University of Technology, China
- **Wyrzykowski, Roman,** Czestochowa University of Technology, Poland
- **Xu, Baomin,** Beijing JiaoTong University, China
- **Zavoral, Filip,** Charles University in Prague, Czech Republic

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 127–131
DOI: 10.15439/2016F523
ACSIS, Vol. 9. ISSN 2300-5963

Innovations from the early user phase on the Jetstream Research Cloud

Richard Knepper, Jeremy Fischer, Craig Stewart, David Hancock and Matthew Link
Pervasive Technology Institute
Indiana University, Bloomington, Indiana 47408
Email: rich@iu.edu

***Abstract*—We describe the Jetstream cyberinfrastructure for research, a purpose-built system with the goal of supporting "long-tail" research by providing a flexible infrastructure that can provide a set cloud services tuned for research applications, whether they be traditional HPC applications, science gateways, or desktop applications. Jetstream offers a library of virtual machines and allows the user to create their own virtual machines in order to provide an open cloud for science that allows both on-demand and persistent instances. The system is currently in early-user mode and a number of users at partner institutions are already creating and using images in the system. This paper details some of the early work being done with the system to create high performance clusters in an on-demand fashion to support scientific work directly as well as serve as capability backend to scientific gateways such as CyVerse and Galaxy.**

I. Introduction

THE JETSTREAM SYSTEM is designed to provide a production cloud resource in support of general science and engineering activities in the eXtreme Digital (XD) ecosystem. While the United States National Science Foundation (NSF) has funded a number of high performance computing (HPC) as well as high throughput computing (HTC) resources, but there remains a significant population of researchers who have computational and data analysis needs that neither HPC nor HTC resources are fit[1]. The NSF has noted [4]the benefits of increasing diversity in the range of cyberinfrastructure (CI) resources available to researchers. Other efforts on the part of the NSF to provide more support for research in the "long tail of science" include the Comet[2] system and the Wrangler data storage and data analytics system[3].

The Jetstream resource and planned activities are well-described in [1]. The current system has been implemented at Indiana University (IU) and Texas Advanced Computing Center (TACC) and is currently in early user mode. In this paper we review the Jetstream architecture and software environment and detail early user experiences with the system, including innovative use of Apache Mesos software for building on-demand cluster resources, using the Atmosphere software suite to serve as a capability service for science gateways, and a "desktop mode" for scientific computing in the field or at sites with limited resources. We discuss some of the challenges of managing a multi-zoned research cloud system in a seamless fashion. We conclude the paper with a discussion of future activities with the service.

II. Jetstream Overview

The Jetstream system is designed to provide general purpose cloud resources for research in a configurable fashion. The system provides on-demand and persistent virtual systems that support a wide range of scientific software in the form of configurable environments.

A. Service Functions

The Jetstream system as designed supports multiple modalities of use in support of scientific research that are currently not provided within the broader cyberinfrastructure ecosystem. These include: self-serve academic cloud services, based on virtual machines images provided by the user or selected from a library; persistent virtual machine systems which support the delivery of science gateways such as the Galaxy [5] life science research gateway; data movement via the Globus Connect and authentication via Globus Auth [6]; facilities for publishing and sharing virtual machine images via IU's persistent digital repository, IUScholarWorks, accessible via Digital Object Identifier (DOI)[7]; and provide virtual desktop services to institutions with limited resources.

Within the cyberinfrastructure ecosystem, the Open Science Grid, the Extreme Science and Engineering Discovery Environment (XSEDE), and other projects provide a broad range of resources for computational support of research: high-performance computing at large scape, high-memory resources, big data, and visualization systems, but to date no system provides a highly configurable virtualized environment with the capabilities described above. The Jetstream system is designed to provide resources for the "long tail" of science [8], who frequently need more access to interactive computational resources than they have locally available, and provide interactive access to a handful of systems as needed in an on-demand fashion, rather than forcing them to work with an allocation system, as in XSEDE, or with a virtual organization, as in the Open Science Grid.

B. Hardware Configuration

The Jetstream system hardware architecture follows commercial cloud offerings in terms of uptime and availability. Two production systems, one at IU and one at TACC, provide a 100Gbps-linked, distributed cluster infrastructure. A third development and testing system resides at the University of Arizona. This configuration offers "zoning" between the two production centers, similar to that offered in Amazon EC2

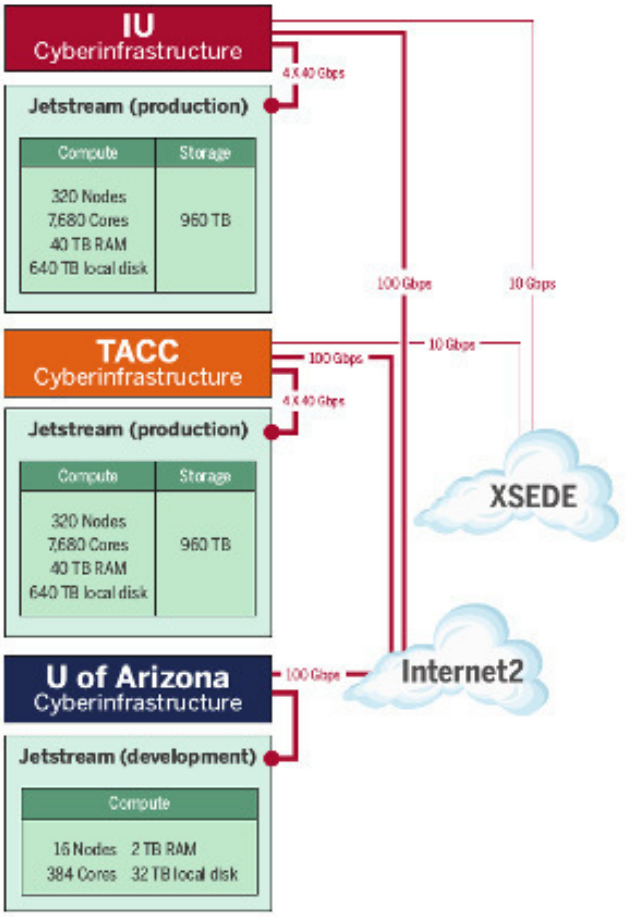


Fig. 1. Jetstream physical sites and network connectivity

resources and elsewhere. A schematic of the Jetstream system is presented in figure 1.

Jetstream provides multiple VM configurations, ranging from "Tiny": 1 CPU, 2GB of memory, 20GB of storage, allowing as many as 46 concurrent virtual machine instances up to "XXL": 44 CPUs, 120GB memory, 480GB storage, with one virtual instance. "Small", "Medium", "Large", and "X-Large" configurations are also offered, with according sizes.

C. Software Architecture

Jetstream utilizes the Atmosphere[9] software stack for presenting a user interface, managing images, provisioning, monitoring and managing cloud infrastructure. Openstack provides host and virtual machine management, as well as virtual machine filesystem storage, with iRODS software providing replication between sites. Authentication is provided by Globus Auth, and user data transfer between the user's desktop and virtual machine filesystems is provided by Globus Transfer. A diagram of the Atmosphere implementation on Jetstream is shown in figure 2.

Atmosphere offers a number of features essential to research computing in a cloud context. Identity management services, networking configuration, and security policies are

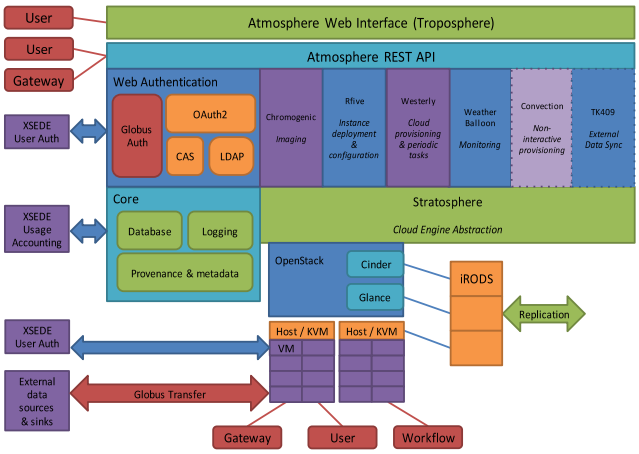


Fig. 2. Atmosphere Architecture in Jetstream

integrated with the software stack. The software provides complete functions for managing virtual machine instances throughout their lifecycle. Finally, data lifecycle management is simplified via the Atmosphere web portal and API. The web portal displays images that can be launched and worked with by the user, shared between users, and new images can be developed and uploaded. Once an image is launched it can be used interactively. A diagram of the Jetstream user interface is presented in figure 3.

III. Modalities of Use

The flexibility of the Jetstream resource offers researchers a broad range of options for computational support of their research. During the initial implementation of Jetstream, the system was placed in "early-user" mode at the beginning of 2016 with the intention of providing the service to the Atmosphere development team at University of Arizona, as

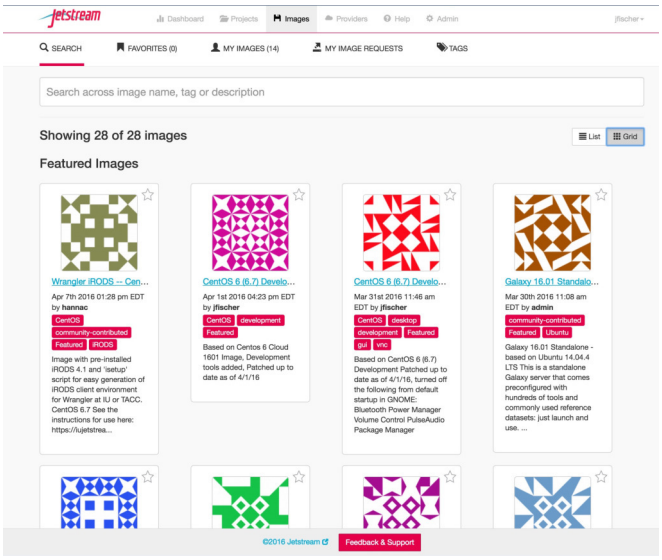


Fig. 3. Jetstream User Interface

well as other scientific users with some familiarity with cloud research in order to test the capabilities of Jetstream and determine what modalities of use might be possible and which ones users may prefer. In this section we describe some of the modalities of use established with these early system users.

A. Research Cloud Capabilities

As a research cloud resource, the Jetstream system must support individual usage of the system directly via interactive sessions with running virtual machine images, but also must support utility computing, particularly in support of science gateway services. In this scenario, users are able to choose workflows and data via a web-based science gateway and workflows are executed in a capability environment via middleware utilties. One of the common software frameworks for supporting science gateway computing is Apache Airavata [10]. In a similar fashion, utility computing resources can be utilized by providing high throughput computing capabilities. Three utility computing modalities for Jetstream have been established in the time since the system was introduced: support for CyVerse; the Galaxy life science gateway; and the ATLAS high energy physics project.

1) CyVerse: One of the initial project goals of the Jetstream system as proposed to the NSF was to improve the availability of compute resources to the CyVerse project[1] (previously known as the iPlant Collaboration). CyVerse was created in order to support life sciences research and improve access to existing cyberinfrastructure, and CyVerse created the Atmosphere cloud service which powers Jetstream. For the CyVerse project, Jetstream provides a set of on-demand system image toolkits, which can be instantiated and run to complete analyses, and then archived or repeated, with the image file available via DOI. Developers at Arizona and IU have created a number of system image toolkits for specific types of research. These toolkits include:

1) a general life science toolkit
2) an R toolkit
3) an astronomy toolkit
4) a data transfer toolkit, with iRODS interface to the Wrangler system at Texas Advanced Computing Center
5) a phylogenetics toolkit

The ability for Jetstream project users to create, save, and share toolkit system images provides a flexible means of collaborating on multi-researcher projects.

2) Galaxy: The Galaxy life science gateway service provides a comprehensive platform for genomic research supporting advanced data management capabilities with an intent to support reproducibility of analyses[5]. The Galaxy web service can utilize multiple different types of resources for analysis, and researchers can either make use of the main Galaxy portal at http://usegalaxy.org or they can set up their own Galaxy servers providing for specific community or lab requirements. Galaxy users at IU have created two means of supporting Galaxy with Jetstream. In the first, the user is able to instantiate a persistent virtual machine image and run a local Galaxy gateway on Jetstream, either completing analyses within the same virtual machine or incorporating other virtual machines in Jetstream to provide additional computational capability. Jetstream VM's are configured as resources for the Galaxy Cloud cyberinfrastructure management tool. Jetstream VM's are with images ready to receive jobs and create a local data caches. The main Galaxy system distributes jobs to VMs via slurm workload manager with pulsar, Galaxy's remote execution service. At this time, about 4 months after Jetstream was initially opened to early use, 7,299 Galaxy main jobs have been run on Jetstream, completing work requests from 758 distinct users.

In the second path, Jetstream system images with Galaxy server deployed are launched within Jetstream and either use their own local resources or submit jobs to Galaxy Cloud. This latter on-demand service provides an easily created and archived Galaxy instance for short to medium term analyses that can be archived in the IU Scholarworks system and retrieved via DOI for the purposes of replication or further analysis.

3) ATLAS: The ATLAS experiment at the Large Hadron Collider utilizes considerable computational resources via the Open Science Grid, many at participating sites. Open Science Grid jobs are well-prepared to take advantage of cloud resources, as Open Science Grid software is able to provide "glide-in" capabilities which allow for jobs to be distributed to resources via a factory. The Jetstream resource is able to support virtual machine images which are a simple base operating system with the software to accept glide-in submissions. ATLAS experiment users can submit jobs to their virtual organization Condor scheduler, which will start the virtual machines on Jetstream via the Atmosphere API. The jobs will be submitted to the virtual machine as glide-in jobs and managed by Open Science Grid monitoring and scheduling resources.

B. Innovative use of Jetstream as research cloud

Early users have also made inroads in optimizing the capability offered by configurable systems for research. Using the Apache Mesos cloud manager system, early users have demonstrated the allocation of virtual system hosts on Jetstream to be managed by Mesos. Mesos provides information on cpu and memory available, and supports a number of frameworks to aid in the management of scaling tools. In this instance, Marathon was used to long-running services in Docker. Within Mesos, users were able to start instances running Docker containers with Unidata IDV, a tool for analyzing and visualizing geoscience data. This early demonstration usage showed that Jetstream could be used as the service resource for a user-defined cluster management framework, instantiating resources and completing work within the temporary cluster, and then releasing those resources to the research cloud and archiving system image information for later use.

IV. Operating a research cloud

A. Functionality tests and operational metrics

In order to establish functionality of the Jetstream system as proposed, a number of functional tests of the openstack environment, atmosphere API, and application software were conducted in order to demonstrate the system's viability as

[1] http://www.cyverse.org

a distributed cloud platform for research activities. A large part of the functional activity tests of Jetstream focus on the set of tasks a user can manage for themselves in order to make use of the architecture, demonstrating a "self-service" system that allows the bulk of activities to take place on the end user's initiative, without requiring the intervention of IT staff. The base set of use cases to articulate this are (assuming that a user has authorization via an allocation and has sufficient understanding of the system):

- Cloud functionality. A sufficiently authorized and knowledgeable user can:
  - authenticate to the Jetstream web interface
  - launch a virtual machine from a library of images on either cluster location
  - quiesce an image running on one cluster, move it to another cluster, and reactivate it
  - create and access a permanent cloud data storage space
  - modify a pre-existing image and store the changed image to the image library on either cluster
- Data management functionality. A sufficiently authorized and knowledgeable user can:
  - move a file from another system into the Jetstream system
  - select a file from Jetstream and move it to another system
  - save a VM image and upload it to IUScholarworks and receive a DOI for the object

The Jetstream system has demonstrated the ability of a user to perform all of these use cases as part of early user mode and acceptance testing for the NSF. Additionally the system has as part of testing activities demonstrated the ability to provide gateway services. These tests consisted of demonstrating that a virtual machine could be run in a continuous fashion, providing access to computational workflows via a standard web interface. In order to test this, the Galaxy and SEAgrid science gateway software were implemented on Jetstream virtual machines with the requirement of running workflows in a comparable amount of time as XSEDE resources. The results gateway services tests are detailed in table I.

TABLE I. GATEWAY FUNCTIONALITY TESTS OF THE JETSTREAM SYSTEM

Test	Success Criteria	Result achieved
Deliver Galaxy gateway	Workflow execution within ≤ 125 % of time required	Workflow completes in 80% of time to run on Stampede
Deliver additional gateway in XSEDE	Correct operation and availability within 2% of overall system availability over 14-day test period	SEAgrid gateway operated continuously for 14 days

Furthermore, the system had a number of operational goals to meet as part of early user stages. These goals include the availability of the system, system capacity, job completion, number of users, number of active VM's, CPU utilization, and number of images published to IUScholarworks. The system met all but two of these goals during early user phase - number of distinct users and active VM's. The outcome of these tests is show in table II.

TABLE II. OPERATIONAL TESTS OF THE JETSTREAM SYSTEM

Metric	Goal	Achieved
System availability (uptime)	95%	100%
% Capacity available	95%	100%
Job completion success	96%	97.7%
Number of distinct users	1,000	327
Use - average number of active VMs	320	290 (1217 peak)
CPU % utilization	6%	4.2% (20.3% peak)
VM images published to IUScholarworks	10/year	6/quarter

B. Challenges of managing the Jetstream system

Along with this early-user utilization of Jetstream, a number of challenges for supporting the research cloud framework have been identified that remain to be addressed. Some of these issues result from the fact that Jetstream incorporates an installation at IU and one at TACC. While this provides a "zoning" function that allows for resources to be available in different location and retain functionality in the event of a large-scale network outage or system-wide maintenance, it also means that these two systems must be kept synchronized in order to ensure coherence for users across both systems. User accounts, managed by LDAP and Globus OAuth, must be synchronized in both places, which has been a manageable process with standard automation tools. Data sync across the clusters represents a more difficult challenge. In order for Jetstream to present a seamless interface, no matter what zone a user accesses, the image library must be synchronized across both cluster, which means that system images up to the XXL size (480GB allocated storage) may need to be copied across clusters. In addition, ensuring that user identification numbers (UIDs) and permissions are the same across both systems. Furthermore, Jetstream engineers are still working on a reliable means of transporting a system across the two clusters, so that a running virtual machine image can be quiesced, relocated, and restarted on a different cluster. This requires both fast data transfer between clusters as well as robust scripts for managing ownership and permissions in both places to ensure that the data can be accessed by the right people.

V. CONCLUSION

We have described some of the early successes and challenges faced by the Jetstream system in its first few months of operation. Jetstream demonstrates efficacy as a research cloud resource, in contrast to projects such as FutureGrid, which are largely used for the exploration of cloud technology and management software, Jetstream projects are capable of supporting recognized research workflows, carried out by researchers, without the aid of grad students or computer science consultants.

The next steps for work on Jetstream are the configuration of images and automation via the OpenStack API to create both persistent and dynamic resources for workflows initiated in the broader cyberinfrastructure (such as through the CyVerse portal or another existing portal) or initiated from gateways running within Jetstream itself. Further work on desktop-like access to Jetstream images will also benefit the Jetstream user base who need short-term interactive resources that can be suspended and revived as necessary.

ACKNOWLEDGMENTS

Implementation of Jetstream is supported by NSF award 1445604. The Indiana University Pervasive Technology Institute was established with support of funding from the Lilly Endowment, Inc., and provided support for this research. CyVerse is supported by NSF award DBI-0735191 and DBI-1265383. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation, or the Lilly Endowment.

REFERENCES

[1] Stewart, Craig A., et al. "Jetstream: A self-provisioned, scalable science and engineering cloud environment." Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure. 2015. ACM: St. Louis, MO, USA. p. 29-37.

[2] Moore, R.L., C. Baru, D. Baxter, G. Fox, A. Majumdar, P. Papadopoulos, W. Pfeiffer, R.S. Sinkovits, S. Strande, M. Tatineni, R.P. Wagner, N. Wilkins-Diehr, M.L. Norman. Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment. 2014. ACM: Atlanta, GA, USA. p. 1-8.

[3] Texas Advanced Computing Center. Wrangler. https://www.tacc.utexas.edu/-/wrangler-data-intensive-system- opens-to-scientists

[4] Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020. 2014. Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017- 2020: Interim Report. Washington, DC. The National Academies Press. 48 pp. http://www.nap.edu/catalog/18972/future-directions-for-nsf-advanced-computing-infrastructure-to-support-us-science-and-engineering-in-2017-2020

[5] Goecks, J, A. Nekrutenko, J. Taylor, and The Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11(8):R86. http://genomebiology.com/2010/11/8/R86

[6] Foster I, Kesselman C. Globus: A metacomputing infrastructure toolkit. International Journal of High Performance Computing Applications. 1997 Jun 1;11(2):115-28.

[7] Bobay J. Institutional Repositories: Why Go There?. Indiana Libraries. 2008 Jan 1;27(1):7-9

[8] Heidorn, P.B. Shedding Light on the Dark Data in the Long Tail of Science. Library Trends. 2008. 57(2), p. 280-299. http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html

[9] Skidmore, E., S.-j. Kim, S. Kuchimanchi, S. Singaram, N. Merchant, and D. Stanzione. iPlant Atmosphere: A Gateway to Cloud Infrastructure for the Plant Sciences. In Proceedings of the 2011 ACM workshop on Gateway computing environments. 2011, ACM: Seattle, Washington, USA. p. 59-64. http://dl.acm.org/citation.cfm?id=2110495

[10] Marru S, Gunathilake L, Herath C, Tangchaisin P, Pierce M, Mattmann C, Singh R, Gunarathne T, Chinthaka E, Gardler R, Slominski A. Apache airavata: a framework for distributed applications and computational workflows. InProceedings of the 2011 ACM workshop on Gateway computing environments 2011 Nov 18 (pp. 21-28). ACM.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 133–140
DOI: 10.15439/2016F52
ACSIS, Vol. 9. ISSN 2300-5963

A Parallel MPI I/O Solution Supported by Byte-addressable Non-volatile RAM Distributed Cache

Artur Malinowski*, Paweł Czarnul*, Piotr Dorożyński*,
Krzysztof Czuryło†, Łukasz Dorau†, Maciej Maciejewski† and Paweł Skowron†
*Gdansk University of Technology, Gdansk, Poland
Email: artur.malinowski@pg.gda.pl, pczarnul@eti.pg.gda.pl, piotr.dorozynski@pg.gda.pl
†Intel Technology Poland Sp. z o.o., Gdansk, Poland
Email: {krzysztof.czurylo, lukasz.dorau, maciej.maciejewski, pawel.skowron}@intel.com

***Abstract*—While many scientific, large-scale applications are data-intensive, fast and efficient I/O operations have become of key importance for HPC environments. We propose an MPI I/O extension based on in-system distributed cache with data located in Non-volatile Random Access Memory (NVRAM) available in each cluster node. The presented architecture makes effective use of NVRAM properties such as persistence and byte-level access behind the MPI I/O API. Another advantage of the proposed solution is making development of a parallel application easy and efficient as a programmer just needs to use the well known MPI I/O data model and API while efficient file access is automatically provided without a need for application level optimizations like avoiding frequent operations on a small data. Results of experiments obtained with three different applications suggest, that the extension significantly reduces file access time, especially for small I/O operations. By locating cache facilities on computing nodes, the extension decreases load of file system servers and makes I/O scalable.**

I. Introduction

SIZES of high performance computing systems are steadily growing. The currently most powerful cluster Tianhe-2 on the TOP500[1] list features 3120000 cores and 1,024,000 GB total memory. It should also be noted that while clusters are larger and larger and potentially allow for higher speed-ups, there are more and more cores and nodes involved in processing and the probability of failure increases. From this point of view, especially in terms of data processed by such applications, there is a need for reliable and large storage solutions that would support execution of such applications. The Message Passing Interface (MPI) standard [1], [2] includes an MPI I/O part that specifies an API for a parallel application to read and write a single file from many processes. Firstly, the API allows both reading/writing data from individual processes or in a collective manner. Secondly, it allows using explicit offsets, individual or shared file pointers.

Within this paper, motivated by possibility of better I/O performance thanks to NVRAM in HPC environments, we propose wrappers over selected MPI I/O API functions using distributed persistent memories in cluster nodes and then compare the performance of the proposed solution with hardware-based simulation of persistent memory to performance of the same MPI application using OrangeFS in a real cluster environment.

The research in the paper was supported by a Grant from Intel Technology Poland.

[1]http://top500.org/system/177999

II. Related Work

While great effort is put into increasing computational power of supercomputers, many data-intensive applications suffer from insufficient I/O operations performance. Speeding up access to storage devices by applying best practices widely proposed in data centers [3], [4], [5] introduces additional overhead for development process, connected with tuning up the application both for each MPI implementation and Parallel File System (PFS). This leads to the conclusion, that a reasonable way would be to apply a generic solution that is suitable for many applications.

Many performance oriented MPI I/O solutions base on the idea of sieving, prefetching and caching data in RAM. ROMIO, a popular MPI I/O implementation, introduces Two-Phase I/O – an algorithm that attempts to merge non-contiguous requests into larger and more contiguous [6]. Tsujita, Y. et al. obtained remarkable improvements by extending Two-Phase I/O even further, by using multiple threads [7]. Other researchers are focused on new MPI I/O implementations [8] or improvements in PFS [9][10]. The extensions of this kind, however, do not consider possibilities offered by emerging hardware technologies.

A significant group of proposed PFS improvement ideas, that could be easily customized to benefit from NVRAM properties, concerns, among other things, cooperative caching. In 1994, Michael D. Dahlin et. al. prepared a survey of different cooperative caching algorithms and showed performance impact of their incorporation into file systems [11]. The described caching strategies are the base for many modern approaches. Our solution differs from others e.g. zFS file system [12] or Novel Distributed Memory File System [13] in management strategy, as we want to avoid central management because of its poor scalability while increasing the number of

cluster nodes. Several papers have proposed extending MPI I/O with cooperative caching algorithms that do not rely on central entity. AHPIOS [14] is an ad-hoc file system, that can be used alternatively to popular PFS implementations. Its main features are tight integration with an MPI application (the client application communicates with a file system using an MPI communicator), minimal configuration and a single instance of global registry with size reduced by keeping metadata minimal. On the other hand, there is no easy method to access created files outside of MPI. Our solution may seem to have a lot in common with another research project, with the mechanism described by Wei-keng Liao et. al. [15], but partially different assumptions (mainly limited amount of memory dedicated to cache storage) led to other technical details. Differences particularly involve splitting cache into pages replaced in our approach by using constant blocks, a complex mechanism of locking unnecessary in our architecture because of request queuing, and a single thread responsible for handling multiple files – our solution serves multiple files simultaneously by taking advantage of a single thread per file.

Another topic, that is also important for parallel, especially long running, applications is checkpointing [16][17]. An application can save its work on a disk or in persistent memory and consequently it can restart from the last known state in case of a failure. I/O bandwidth is an important factor in reducing execution time. In 2013, Rajachandrasekar et. al. proposed CRUISE – in memory file system that speeds up checkpointing [18]. In this system, each write request data is initially stored in a pre-allocated persistent memory region, and after flushed to a PFS or a local file system asynchronously. Performance results presented by the authors were really promising, nevertheless, checkpointing has different characteristics than operating on a file while performing computations, therefore it cannot be applied to our solution. In many cases, the main disparity is connected to a single process operating on a single file, what reduces complexity of routines responsible for simultaneous accesses. Other differences in assumptions, that make checkpointing optimizations inadequate in our solution, are: strong spatial locality of requests (accessed data is rather continuous), usually central management of a checkpoint and focusing much more on optimizations of output operations.

I/O operations are strongly related to storage hardware. In 2009, Mark H. Kryder and Chang Soo Kim evaluated several memory technologies that are expected to be an alternative for hard disk drives (HDD) in 2020 [19]. Some investigated solutions have interesting properties such as non-volatility and random access (NVRAM), fast read/write access time and high density (which affects final capacity of a device) while the price could be still reasonable compared to HDD. In 2015, Intel Corp. and Micron Technology unveiled 3D XPoint – non-volatile memory technology expected to be up to 1,000 times faster than NAND, 10 times denser than DRAM with latency of tens of nanoseconds and possible to be used as system memory [20], [21]. With declared relatively low price and expected market release in 2016 [22], 3D XPoint announcements show, that NVRAM has a potential to be a true alternative for existing storage technologies soon.

Many MPI I/O extensions benefit from particular storage hardware properties. Shuibing He et al. implemented Solid State Drive (SSD) cache that improved throughput of PFS [23][24], however, block data access and long latency of SSDs cause, that the solution is not able to benefit from all properties available in NVRAM. Evaluation of NVRAM role in data-intensive scientific applications was presented by Dong Li et al. [25] and – independently – Brian Van Essen et al [26], but papers are based on single node analysis and are focused rather on extending system memory (heap, stack, global data segments) than speeding up I/O operations in a distributed environment. Active NVRAM for I/O staging proposed by S. Kannan et al. [27], [28] benefits from NVRAM located within each computing node speeding access to PFS up. While this solution could be useful in the case considered in this paper, it does not fulfill our requirement of minimal application modification understood as keeping the proposed extension compatible with MPI I/O API. Moreover, the presented experimental results suggest, that this solution is not beneficial for small data sizes – we assume, that our extension is convenient for developers in the way it allows to access even very small data efficiently.

III. Motivations

In view of the existing solutions and recent developments in the area of non-volatile RAM, new solutions could be proposed for parallel applications that could potentially increase both performance and ease of development of applications processing potentially large data sets. Specifically:

1) Performance. It is possible to use a collection of distributed persistent memories in cluster nodes as an additional layer of cache between an underlying file system and an application. It can serve as an intermediate layer able to store large data sets (larger than in the combined RAM of cluster nodes) with persistence and possibility to recover from persistent memory should a failure occur. Thanks to the relatively low latency of persistent memory, this should allow a solution with better performance than traditional file systems, especially if an application would perform reads or writes to far away spaced locations in a file.
2) Ease of development/programming/data model. Such a solution with a proper API could in fact be regarded as a shared (distributed in the underlying implementation), large memory with persistence and, what is important, byte level access which is a property of persistent memory. While block level access would still yield better performance, even accesses using small blocks or even bytes could yield much better performance than traditionally used file systems for parallel applications.

IV. Proposed Solution

A. Design

This solution is not another MPI I/O implementation – the contribution of the paper is a set of wrappers over MPI I/O

API functions that creates in-app distributed cache between application and particular MPI I/O implementation. Source code is written in C using MPI, POSIX Threads API and the libpmem library responsible for low level NVRAM memory support [29]. Wrappers incorporate NVRAM usage behind the MPI I/O API.

The extension requires a specific architecture as shown in Fig. 1. We assume a cluster with interconnected nodes each of which allow running processes of an application in parallel. Each node must be equipped with its own NVRAM storage, where the cache data is stored. All computing nodes, as in the regular MPI I/O, must have access to a remote file using a distributed file system.

A considered file to be accessed using the MPI I/O API is split into n continuous parts, where n is the number of nodes. Each part is managed by a single, independent cache manager running on a dedicated thread as presented in Fig. 2. Cache managers are mainly responsible for:

- prefetching whole data part – prefetching all of the required data is possible due to the assumption that the size of a file is limited to the sum of all NVRAM capacities in a cluster;
- synchronizing data between cache and file system – occurs only when the file is being opened, closed or the synchronization is called explicitly by the application;
- serving read/write requests from all of the processes in application. Each process is able to determine which cache manager it should contact. Same file locations can be accessed by all processes, which, in case of write requests may require synchronization at the application level.

The proposed solution has no central management. Each process knows exactly which cache manager holds the data, so no additional entity, like dispatcher, is required. Because each cache part is continuous – instead of being split into blocks – metadata is kept to a minimum. The cache manager does not perform any staging optimization – each request is served as fast as possible by making use of NVRAM byte-addressing and low latency compared to HDD or SSD. Processing without a central entity allows to avoid potential bottlenecks, while simplifying data access, and, rather than introducing smart but costly data management, it saves CPU time, reduces latency and makes the solution more independent of specific data patterns.

Although the proposed extension is not a file system, it can serve multiple files simultaneously. Each opened file has its own part of allocated NVRAM, a dedicated thread for a cache manager and an MPI communicator. Required metadata (e.g. file path, file size, communicator handler) is stored within a separate file handler that is returned by a call to the `MPI_File_open` function.

A natural advantage of using NVRAM as a cache storage is its persistent character which is directly linked to the possibility to recover data after failure, but guaranteeing full data consistency in a distributed environment requires further investigation and will be the subject of our next research.

B. Target applications

As presented in Fig. 3a, the solution should be most beneficial in applications that access small data chunks (gain from byte addressing) from spread file locations (no drawback from omitting staging phase). Improved performance is a result of fast read and write accesses, but prefetching a large amount of data in the beginning and the need for writing the whole cache back at the time of closing file introduce overhead associated with initialization and de-initialization. This leads to the conclusion, that in order to perform better than the regular MPI I/O, an application has to access data frequently. As shown in Fig. 3b, it could be achieved either in very data-intensive applications, or in long running applications. However, many scientific applications meet these criteria.

Introducing the file size limitation is not an issue, because NVRAM capacity multiplied by number of nodes in modern clusters is expected to be enough for handling files of the sizes comparable to the SSD based solutions. Our extension is also scalable, so it is expected to perform well while increasing the number of processes or nodes. On the other hand, it should be kept in mind that the total number of processes in an application results in a certain number of processed served, on average, by each cache manager in a cluster node.

C. Implementation

Making use of proposed extension in an MPI application requires two minor changes in source code. The first one is including `file_io_pmem_wrappers.h` header that allows to transform each native MPI I/O function call into its NVRAM cache counterpart. Due to compatibility of function signatures, calls do not need modifications.

Configuration of the solution is prepared with `MPI_Info` parameters passed to `MPI_File_open`. A minimal configuration requires only one parameter, `pmem_path`, that points to an NVRAM device mounted in a local file system. `MPI_Info` parameters unrecognized by MPI I/O are ignored, so the cache could be switched on and off using an include directive.

The extension spawns additional POSIX threads (cache managers) that use MPI to communicate with the application. Therefore, initialization with `MPI_Init_thread` and `MPI_THREAD_MULTIPLE` support are needed. Algorithm 1 shows the idea behind implementation of the cache manager thread. The listing contains all MPI and NVRAM cache related calls. The thread is created within `MPI_File_open`.

An object, that represents a file opened with MPI I/O, is called a file handler. In our solution, the handler is considered as a pointer to `MPI_File_pmem_structure`. Although the object-oriented programming (OOP) paradigm is not natively provided in the C programming language, we used it by incorporating into structure both data together with a set of pointers to functions. Data stored in the structure includes:

- information about the file e.g. name and size,
- handlers responsible for communication with the cache (MPI communicator),
- cache metadata (number of cache nodes, file offsets handled by each cache manager),

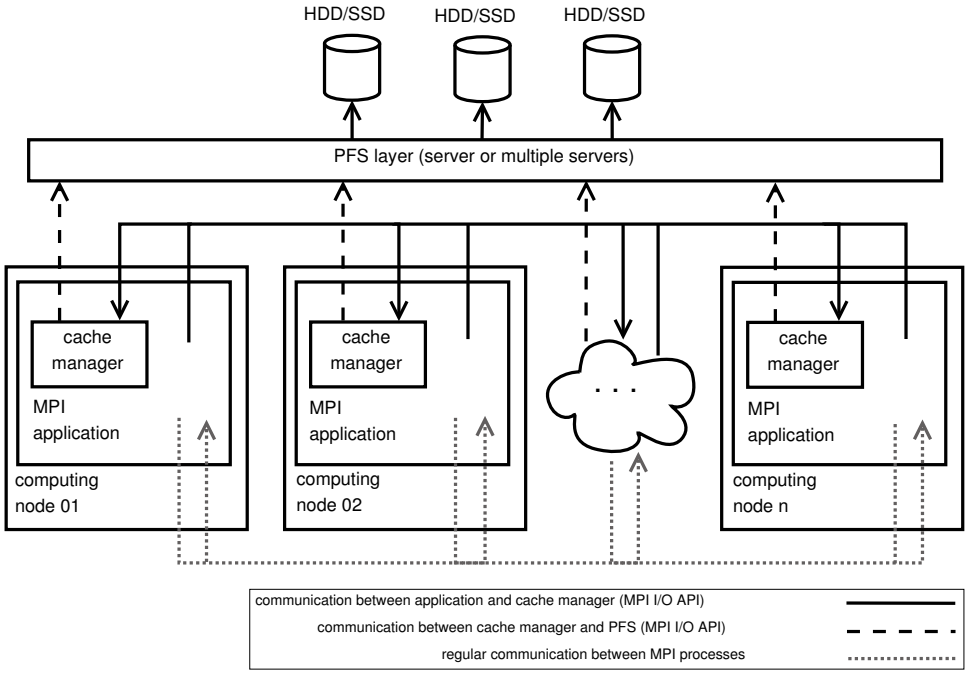


Fig. 1: Architecture of the multi-node system that utilizes the proposed solution. MPI processes are not included in the diagram.

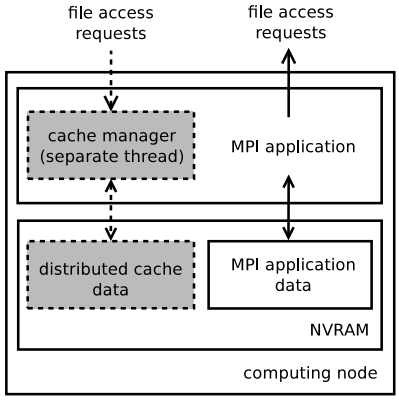


Fig. 2: Architecture of proposed solution within a single node. Gray components and dashed connections are transparent to MPI application developer

- some additional parameters e.g. related with failure recovery.

A set of pointers to functions contain a single counterpart function for each MPI I/O routine. Combining a file with functions allows for choosing different strategies for different files which could be potentially beneficial when extending the method further.

Source code of the solution released under the BSD license, software documentation and examples are available on GitHub platform [2].

[2]https://github.com/pmem/mpi-pmem-ext

V. Experiments

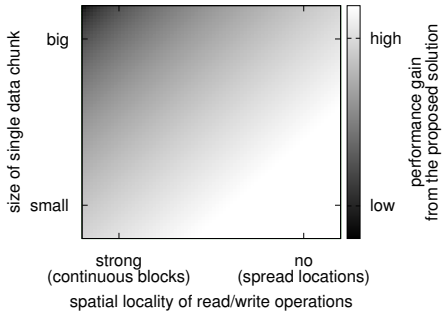
A. Testbed Environment

All of the tests were performed on an eight-node cluster, each node equipped with two Intel® Xeon® E5-4620 CPUs, as well as 32GB of RAM, storage on SSD + HDD together with both 10 Gigabit Ethernet (10GbE) and Infiniband connections.

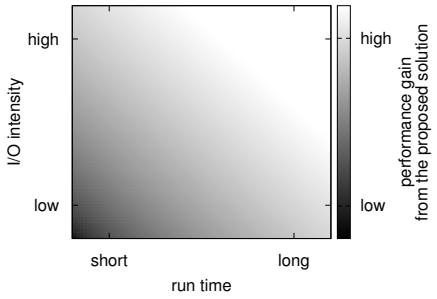
A single node was responsible for application execution, gathering the results and hosting a PFS server, the other 7 nodes for parallel application execution. OrangeFS 2.8.7 (former: PVFS2)[3] compiled with Infiniband support was chosen as PFS, because of relatively good performance [30]. MPICH 3.1.4 with ROMIO[4] was installed on seven computing nodes as an MPI IO implementation. OrangeFS stored both data and metadata on SSD. Nodes were communicating with each other using 10GbE, all of the Infiniband bandwidth was used to provide fast file access. In most experiments each computing node ran 15 processes – with 16 physical cores on a single node it left a spare core for a PFS thread.

RAM in each of seven computing nodes was split into two parts: regular system memory (15GB) and storage for NVRAM simulation (17GB). Amount of NVRAM memory does not influence performance, because the cache manager would use only as much NVRAM, as the size of its cache part. The NVRAM simulation part was visible in the operating system as an ext4 file partition using The Persistent Memory Driver and ext4 Direct Access (DAX)[5]. DAX provided a way

[3]http://www.orangefs.org/
[4]https://www.mpich.org/
[5]https://www.kernel.org/doc/Documentation/filesystems/dax.txt



(a)



(b)

Fig. 3: Plots present properties of the applications, that potentially would benefit most from the proposed solution

to use NVRAM through file system omitting paging, caching etc., so we noted performance comparable to RAM. To obtain expected NVRAM properties, we used a hardware simulator configured with three parameters:

- latency – additional latency to access the data (default: 600ns),
- commit latency – time required to ensure that saved data is flushed on device (default: 2000ns),
- bandwidth (default: 9.5GB/s).

If it is not explicitly stated in the experiment description, the simulator was configured with default values.

B. Results

1) Rompio benchmark and tests: Rompio[6], software developed at Los Alamos National Laboratory, is a file I/O performance benchmark with MPI support. Rompio was chosen because it is able not only to provide a final bandwidth, but also intermediate values (i.e. time of opening or closing a file) useful in performance tuning.

Fig. 4 shows execution times of read and write operations separately, both for NVRAM cache based extension and regular MPI I/O. In this test, the proposed extension is better for small data chunks (up to 1024B), while the regular MPI implementation has better results for larger data.

2) Discussion: The reason for execution time growth that occurs in this case is related to the specific design of the benchmark i.e. the size of a file increases linearly with the size of data chunk, the number of operations and the number of processes. A significant part of execution time of the code using this solution is consumed on opening and closing the file, as it prefetches data into cache, so this extension benefits mostly for long-running applications with many read and accesses on an open file. While the benchmark is very configurable, it does not allow to use a fixed size of a file.

Bandwidth calculated with values that neglect time consumed by opening and closing the file is presented in Fig. 5 which shows much better values for the NVRAM based

[6]http://www.osti.gov/scitech/biblio/1231008-rompio

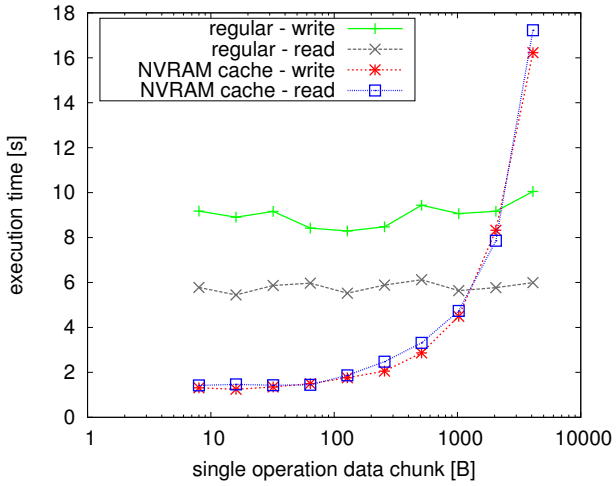


Fig. 4: Rompio benchmark execution time results

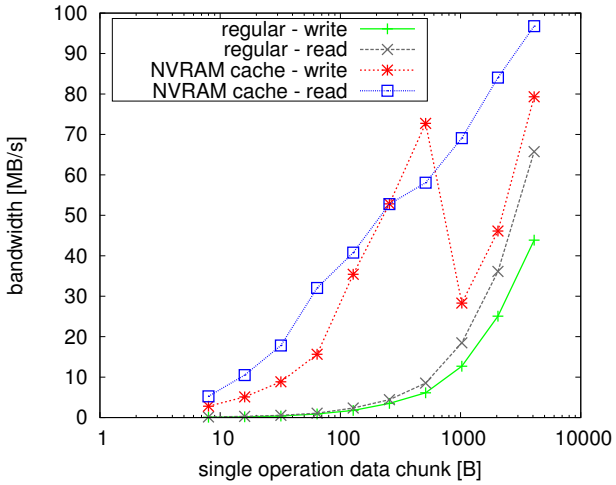


Fig. 5: Rompio benchmark bandwidth results

Algorithm 1 Cache manager routine

```
init_cache(); // allocate NVRAM memory

// prefetch part of file, that
// cache manager is responsible for
MPI_File_read_at();

while (true) {

  MPI_Probe();
  switch (probe_status.MPI_TAG) {

    case READ_AT_REQUEST_TAG:
      MPI_Recv(); // get read request
      read_from_cache();
      MPI_Send(); // send bytes from cache
      break;

    case WRITE_AT_REQUEST_TAG:
      MPI_Recv(); // get write request
      write_into_cache();
      cache_flush(); // flush into NVRAM
      break;

    case SYNC_TAG:
      MPI_Recv(); // get sync request
      MPI_File_write_at(); // flush
                           // into PFS
      break;

    case SHUTDOWN_TAG:
      MPI_Recv(); // get shutdown request
      MPI_File_write_at(); // flush
                           // into PFS
      deinit_cache();
      return;

    // another cases
  }
}
```

solution. As a consequence, in order to achieve a better overall execution time compared to a standard solution, as shown in Fig. 4, the NVRAM based proposed solution needs a high ratio of the time spent on read/write operations compared to the initialization/finalization time spent on open/close operations. The large bandwidth drop between data chunk of 512B and 1024B is caused by inefficiency of asynchronous writing. In the proposed extension, small write requests end immediately after being submitted and then the cache manager is performing an actual writing procedure. However, for constant requests frequency, writing bigger chunks consumes more time, so consecutive requests have to queue.

3) 2D map search and tests: 2D map search is a geometric SPMD type application for searching throughout a 2D map stored in a file. The goal of this application is as follows: search throughout the map for pixels that meet certain criteria and – after a pixel/object meeting a criterion has been found – a part of its immediate surrounding in a selected direction is searched up to a predefined radius or until a given number of pixels meeting another criterion is found. An exemplary application of such an algorithm may be searching for spreading of pollution in farmlands with wind blowing in a certain direction. The application can read the data byte by byte (naive approach, but fastest in development) or use block reading with blocks of a predefined size.

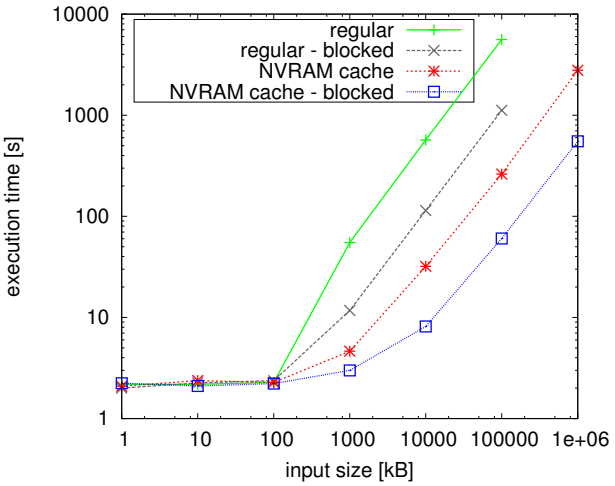


Fig. 6: 2D map search results according to input size (105 processes, 512B block size)

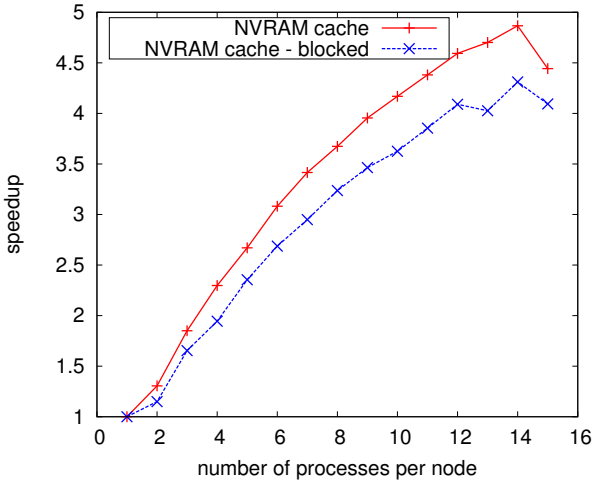


Fig. 7: 2D map search speedup according to number of processes per each of seven nodes (map size: 100MB)

Fig. 6 shows, that for this application the proposed extension performed better than regular MPI I/O, when the size of a file

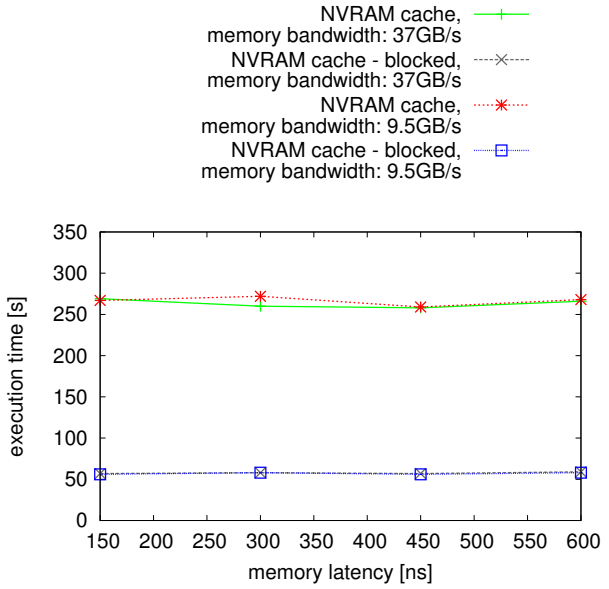


Fig. 8: 2D map search execution time with different NVRAM simulation platform configurations (map size: 100MB)

was greater than 100 KBs (execution time with smaller maps is determined by operations unrelated to I/O). Fig. 7 presents speedup according to the number of processes respectively. 2D map search does not perform any time consuming calculations, execution time is mainly based on I/O operations.

4) Discussion: With regular MPI I/O, the application does not scale because from the PFS perspective, the number and sizes of requests are constant. On the other hand, the proposed extension is scalable – each additional node reduces average load for a single node. Different NVRAM simulation platform configurations do not influence the performance, which is shown in Fig. 8. Taking into consideration file size 100MB, the number of nodes equal to 7, file size per node equal to $\frac{100MB}{7} \approx 14MB$, potential difference between latencies 450ns, for byte level access we can compute the overhead of $450ns \cdot \frac{14MB}{1B} \approx 6.3s$ which constitutes 2.4% of execution time. For 512B blocks we can compute a theoretical overhead of $450ns \cdot \frac{14MB}{512B} \approx 12ms$ while for reference for SSD with 512B block, $0.1ms \cdot \frac{14MB}{512B} \approx 2.7s$. In test runs, we did observe differences in times varying from run to run, in the order of this overhead, coming most likely from file system operations and consequently such overhead is not exposed in the chart.

5) Random walk microbenchmark and tests: A third group of experiments was performed with an application not as data-intensive as Rompio or 2D map search, created in order to check whether the solution could be useful in programs where file operations consume less amount of time compared to the application running time. This microbenchmark is a constrained version of a random walk algorithm. In each step a data chunk is read, the application performs some selected computations (about a million iterations of Collatz conjecture), and the chunk is written back.

6) Discussion: Results presented in Fig. 9 show, that if the ratio of read/write to open/close operations is relatively high, the solution performs better than regular MPI I/O. The dependence shows, that the potential target of the extension is not only a set of data-intensive applications that operate on relatively small data chunks. Programs of a less data-intensive character and operating on bigger data chunks can also benefit from the solution if only they run long enough to compensate the overhead for initialization and de-initialization of NVRAM cache.

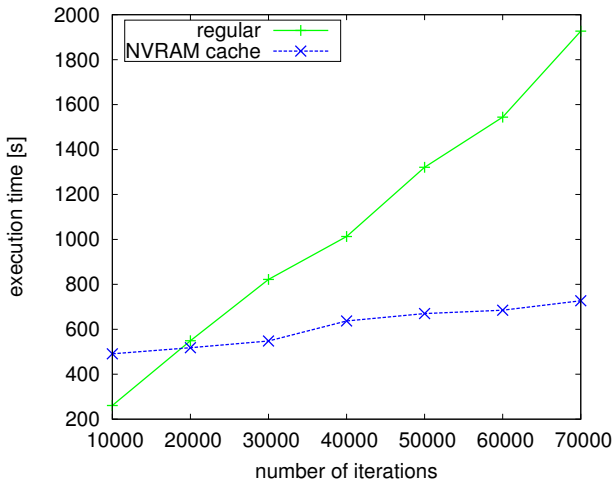


Fig. 9: Random walk microbenchmark execution time results (input file size: 10GB)

VI. Conclusions and Future Work

In this paper, we presented a new parallel MPI I/O solution implemented by our group, including implementation and tests, supported by byte-addressable non-volatile RAM distributed cache. We demonstrated improvements of I/O operations' performance in a cluster environment using NVRAM. We proposed an MPI I/O extension based on distributed cache in NVRAM, not only for improvement of performance, but also to make application the development process easier by allowing accessing small data chunks efficiently using the MPI I/O data model and API. The solution was tested on a cluster equipped with a hardware NVRAM simulator using three different applications: an MPI I/O benchmark, searching throughout a 2D map stored in a file and a microbenchmark based on a random walk algorithm combined with Collatz conjecture. The results confirmed, that in tested applications in a cluster with hardware simulated NVRAM the proposed solution significantly improves performance of small I/O operations, compared to a standard MPI implementation on a typical cluster without NVRAM.

In the nearest future we plan to extend the method further and test selected optimizations. The next step is to test the solution with more applications. At the time of writing, a simulation of tornadoes moving across an area is being prepared.

We also plan on using this approach for parallelization of processing of many images extending the work performed in [31] for parallelization of image processing within GIMP using an NVRAM-assisted MPI based solution. Although the proposed distributed cache is always persistent and can be recreated after a failure, additional set of tests, performance tuning and further research of data consistency are also planned.

REFERENCES

[1] Message Passing Interface Forum, "MPI: A Message-Passing Interface Standard Version 3.1," June 2015, http://www.mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf.

[2] W. Gropp, T. Hoefler, R. Thakur, and E. Lusk, *Using Advanced MPI: Modern Features of the Message-Passing Interface (Scientific and Engineering Computation)*. The MIT Press, 2014, ISBN 978-0262527637.

[3] P. Wautelet, "Best practices for parallel IO and MPI-IO hints," March 2015, http://www.idris.fr/media/docs/docu/idris/idris_patc_hints_proj.pdf.

[4] B. Hadri, "Introduction to Parallel I/O," October 2011, https://www.olcf.ornl.gov/wp-content/uploads/2011/10/Fall_IO.pdf.

[5] N. H.-E. C. Program, "Lustre Best Practices," August 2015, http://www.nas.nasa.gov/hecc/support/kb/lustre-best-practices_226.html.

[6] R. Thakur, W. Gropp, and E. Lusk, "Data sieving and collective I/O in romio," *Frontiers '99 - Seventh Symposium On Frontiers Massively Parallel Computation, Proc.*, pp. 182–189, 1999. doi: 10.1109/FMPC.1999.750599. [Online]. Available: http://dx.doi.org/10.1109/FMPC.1999.750599

[7] Y. Tsujita, K. Yoshinaga, A. Hori, M. Sato, M. Namiki, and Y. Ishikawa, "Multithreaded Two-Phase I/O: Improving Collective MPI-IO Performance on a Lustre File system," *2014 22nd Euromicro Int. Conference On Parallel, Distributed, Network-based Processing (pdp 2014)*, pp. 232–235, 2014. doi: 10.1109/PDP.2014.46. [Online]. Available: http://dx.doi.org/10.1109/PDP.2014.46

[8] A. Hori, K. Yamamoto, and Y. Ishikawa, "Catwalk-ROMIO: A Cost-Effective MPI-IO," *2011 IEEE 17th Int. Conference On Parallel Distributed Systems (icpads)*, pp. 120–126, 2011. doi: 10.1109/ICPADS.2011.40. [Online]. Available: http://dx.doi.org/10.1109/ICPADS.2011.40

[9] F. Wang, Y. Chen, S. Li, F. Yang, and B. Xiao, "The design of data storage system based on lustre for {EAST}," *Fusion Engineering and Design*, pp. –, 2016. doi: 10.1016/j.fusengdes.2016.04.002. [Online]. Available: http://dx.doi.org/10.1016/j.fusengdes.2016.04.002

[10] S. A. Wright, S. D. Hammond, S. J. Pennycook, I. Miller, J. A. Herdman, and S. A. Jarvis, "Ldplfs: Improving I/O Performance Without Application modification," *2012 IEEE 26th Int. Parallel Distributed Processing Symposium Workshops & Phd Forum (ipdpsw)*, pp. 1352–1359, 2012. doi: 10.1109/IPDPSW.2012.172. [Online]. Available: http://dx.doi.org/10.1109/IPDPSW.2012.172

[11] M. D. Dahlin, R. Y. Wang, T. E. Anderson, and D. A. Patterson, "Cooperative caching: Using remote client memory to improve file system performance," in *Proceedings of the 1st USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI '94, 1994. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267638.1267657

[12] A. Teperman and A. Weit, "Improving Performance of Distributed File System Using OSDs and Cooperative Cache," *IBM Haifa Labs*, 2004.

[13] U. Karnani, R. Kalmady, P. Chand, A. Bhattacharjee, and B. S. Jagadeesh, "Design and Implementation of a Novel Distributed Memory File System," ser. Communications in Computer and Information Science, vol. 133, no. III, 2011. doi: 10.1007/978-3-642-17881-8_14 pp. 139–148, 1st International Conference on Computer Science and Information Technology, 2011, India. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17881-8_14

[14] F. Isailå, J. G. Blas, J. Carretero, W.-k. Liao, and A. Choudhary, "AHPIOS: An MPI-Based Ad Hoc Parallel I/O System," in *Parallel and Distributed Systems, 2008. ICPADS'08. 14th IEEE International Conference on*. IEEE, 2008. doi: 10.1109/ICPADS.2008.50 pp. 253–260. [Online]. Available: http://dx.doi.org/10.1109/ICPADS.2008.50

[15] W.-K. Liao, K. Coloma, A. Choudhary, and L. Ward, "Cooperative Client-Side File Caching for MPI Applications," *Int. J. High Perform. Comput. Appl.*, vol. 21, no. 2, pp. 144–154, May 2007. doi: 10.1177/1094342007077857. [Online]. Available: http://dx.doi.org/10.1177/1094342007077857

[16] P. Czarnul and M. Frączak, *Recent Advances in Parallel Virtual Machine and Message Passing Interface: 12th European PVM/MPI Users' Group Meeting Sorrento, Italy, September 18-21, 2005. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, ch. New User-Guided and ckpt-Based Checkpointing Libraries for Parallel MPI Applications,, pp. 351–358. ISBN 978-3-540-31943-6. [Online]. Available: http://dx.doi.org/10.1007/11557265_46

[17] P. Dorożyński, P. Czarnul, A. Malinowski, K. Czuryło, Ł. Dorau, M. Maciejewski, and P. Skowron, "Checkpointing of Parallel MPI Applications using MPI One-sided API with Support for Byte-addressable Non-volatile RAM," *Procedia Computer Science*, vol. 80, pp. 30 – 40, 2016. doi: 10.1016/j.procs.2016.05.295 International Conference on Computational Science 2016, June 2016, USA. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2016.05.295

[18] R. Rajachandrasekar, A. Moody, K. Mohror, and D. Panda, "A 1PB/s File System to Checkpoint Three Million MPI Tasks," June 2013.

[19] M. H. Kryder and C. S. Kim, "After Hard Drives — What Comes Next?" *Magnetics, IEEE Transactions on*, vol. 45, no. 10, pp. 3406–3413, Oct 2009. doi: 10.1109/TMAG.2009.2024163. [Online]. Available: http://dx.doi.org/10.1109/TMAG.2009.2024163

[20] Intel Corporation, "Intel and Micron Produce Breakthrough Memory Technology," July 2015, http://newsroom.intel.com/community/intel_newsroom/blog/2015/07/28/intel-and-micron-produce-breakthrough-memory-technology.

[21] ——, "3D XPoint Technology Revolutionizes Storage Memory," July 2015, http://www.intel.com/content/www/us/en/architecture-and-technology/3d-xpoint-technology-animation.html.

[22] ——, "Introducing Breakthrough Memory Technology," July 2015, http://www.intel.com/content/www/us/en/architecture-and-technology/non-volatile-memory.html.

[23] S. He, X.-H. Sun, and B. Feng, "S4d-cache: Smart Selective SSD Cache for Parallel I/O systems," *2014 Ieee 34th Int. Conference On Distributed Computing Systems (icdcs 2014)*, pp. 514–523, 2014. doi: 10.1109/ICDCS.2014.59. [Online]. Available: http://dx.doi.org/10.1109/ICDCS.2014.59

[24] S. He, Y. Wang, and X.-H. Sun, "Improving Performance of Parallel I/O Systems through Selective and Layout-Aware SSD Cache," *IEEE Transactions on Parallel and Distributed Systems*, 2016. doi: 10.1109/TPDS.2016.2521363. [Online]. Available: http://dx.doi.org/10.1109/TPDS.2016.2521363

[25] D. Li, J. S. Vetter, G. Marin, C. McCurdy, C. Cira, Z. Liu, and W. Yu, "Identifying Opportunities for Byte-Addressable Non-Volatile Memory in Extreme-ScaleScientific applications," *2012 Ieee 26th Int. Parallel Distributed Processing Symposium (ipdps)*, pp. 945–956, 2012. doi: 10.1109/IPDPS.2012.89. [Online]. Available: http://dx.doi.org/10.1109/IPDPS.2012.89

[26] B. V. Essen, R. Pearce, S. Ames, and M. Gokhale, "On the role of NVRAM in data-intensive architectures: an evaluation," *2012 Ieee 26th Int. Parallel Distributed Processing Symposium (ipdps)*, pp. 703–714, 2012. doi: 10.1109/IPDPS.2012.69. [Online]. Available: http://dx.doi.org/10.1109/IPDPS.2012.69

[27] S. Kannan, A. Gavrilovska, K. Schwan, D. Milojicic, and V. Talwar, "Using Active NVRAM for I/O Staging," in *Proceedings of the 2Nd International Workshop on Petascal Data Analytics: Challenges and Opportunities*, ser. PDAC '11. ACM, 2011. doi: 10.1145/2110205.2110209. ISBN 978-1-4503-1130-4 pp. 15–22. [Online]. Available: http://dx.doi.org/10.1145/2110205.2110209

[28] S. Kannan, D. Milojicic, V. Talwar, A. Gavrilovska, K. Schwan, and H. Abbasi, "Using Active NVRAM for Cloud I/O," in *Proceedings of the 2011 Sixth Open Cirrus Summit*, ser. OCS '11. IEEE Computer Society, 2011. doi: 10.1109/OCS.2011.12. ISBN 978-0-7695-4650-6 pp. 32–36. [Online]. Available: http://dx.doi.org/10.1109/OCS.2011.12

[29] NVM Library team at Intel Corporation, led by Andy Rudoff, "pmem.io Persistent Memory Programming," http://pmem.io/nvml/libpmem/.

[30] J. M. Kunkel and T. Ludwig, "Performance evaluation of the PVFS2 architecture," *15th Euromicro International Conference On Parallel, Distributed And Network-based Processing, Proceedings*, pp. 509–516, 2007.

[31] P. Czarnul, A. Ciereszko, and M. Fraczak, "Towards efficient parallel image processing on cluster grids using gimp," in *Computational Science - ICCS 2004*, ser. Lecture Notes in Computer Science, M. Bubak, G. van Albada, P. Sloot, and J. Dongarra, Eds. Springer Berlin Heidelberg, 2004, vol. 3037, pp. 451–458. ISBN 978-3-540-22115-9. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-24687-9_57

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 141–145
DOI: 10.15439/2016F327
ACSIS, Vol. 9. ISSN 2300-5963

Energy-efficient FPGA Implementation of the k-Nearest Neighbors Algorithm Using OpenCL

Fahad Bin Muslim, Alexandros Demian, Liang Ma, Luciano Lavagno
Department of Electronics and Telecommunication
Politecnico di Torino, ITALY

Affaq Qamar
Department of Electrical Engineering
Abasyn University, Peshawar Pakistan

***Abstract*—Modern SoCs are getting increasingly heterogeneous with a combination of multi-core architectures and hardware accelerators to speed up the execution of compute-intensive tasks at considerably lower power consumption. Modern FPGAs, due to their reasonable execution speed and comparatively lower power consumption, are strong competitors to the traditional GPU based accelerators. High-level Synthesis (HLS) simplifies FPGA programming by allowing designers to program FPGAs in several high-level languages e.g. C/C++, OpenCL and SystemC.**

This work focuses on using an HLS based methodology to implement a widely used classification algorithm i.e. k-nearest neighbor on an FPGA based platform directly from its OpenCL code. Multiple fairly different implementations of the algorithm are considered and their performance on FPGA and GPU is compared. It is concluded that the FPGA generally proves to be more power efficient as compared to the GPU. Furthermore, using an FPGA-specific OpenCL coding style and providing appropriate HLS directives can yield an FPGA implementation comparable to a GPU also in terms of execution time.

***Keywords*—kNN; FPGA; High-Level Synthesis; Hardware Acceleration; low-power low-energy computation; Parallel Computing; OpenCL.**

I. Introduction

The ever increasing requirement for electronic devices to perform a variety of compute intensive operations has resulted in the evolution of advanced system-on-chip (SoC) designs with heterogeneous system architectures. These heterogeneous systems are essentially multi-core systems offering a substantial gain in performance not only by utilizing additional cores but also by embedding specialized hardware accelerators e.g. Graphics Processing Units (GPUs) and field programmable gate arrays (FPGAs) for accelerating various compute intensive parts of complex applications. A simplified overview of such a system is shown in Fig. 1.

These systems offer substantial gains in execution time as well as energy efficiency [1]. Modern high performance computing (HPC) systems thus rely on such heterogeneous systems consisting of traditional processors for performing the sequential tasks and FPGAs used as accelerators performing tasks concurrently. Modern FPGAs have the ability to provide sufficient processing speed while consuming a fraction of the power consumed by high-end GPUs [2]. This is why several big data companies such as Microsoft, Baidu are exploring FPGA devices as accelerators rather than GPUs [3, 4].

The major limitation while considering such system architectures is the complexity to program the FPGAs which traditionally requires a considerable expertise in register transfer level (RTL) design. This issue is addressed by an approach called high-level synthesis (HLS), which tends to reduce both the verification and design time and effort for an FPGA based application by allowing the designers to program in several higher level languages such as C, C++ and SystemC.

A very promising parallel programming language which is built upon C/C++ and can be used to program an FPGA is the Open Computing Language (OpenCL). The fact that OpenCL is based upon C/C++, makes porting a program from C/C++ to OpenCL quite easy [5]. OpenCL is a programming standard developed by the Khronos group to develop applications being executed on heterogeneous platforms. OpenCL due to its portability holds an edge over the very similar Compute Unified Device Architecture (CUDA) programming framework, which can be used to program NVIDIA GPUs only. Though OpenCL is device portable, yet it does not offer performance portability across multiple devices. OpenCL program support on Xilinx FPGA devices is provided by SDAccel™, which is a Xilinx development environment for synthesizing OpenCL kernels to be executed on Xilinx FPGA devices [6]. The Xilinx OpenCL high-level synthesis tool, namely Vivado HLS, has been used in this work to implement the k-nearest neighbor (kNN) algorithm onto Xilinx FPGAs.

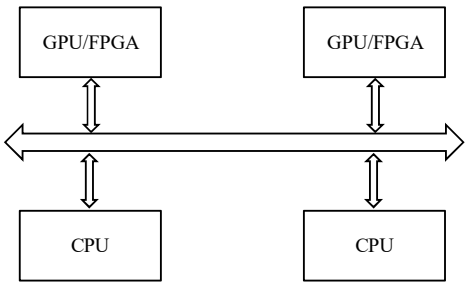


Fig. 1. Typical Heterogeneous System Architecture

The kNN algorithm is used to find the *k* nearest neighbors of a specific point among a set of unstructured data points. It is widely used in a diverse range of domains and applications such as pattern recognition, machine learning, computer vision and coding theory to name a few. The algorithm

unfortunately has a significantly large computation cost, since typically training data sets are very large [7]. The algorithm though is highly parallelizable and can be accelerated considerably by exploiting the inherent parallelism of an FPGA device.

This work starts from a parallel implementation of the kNN algorithm from the Rodinia library [8]. Two different implementations of the algorithm are considered and their time, energy/power and cost performance are compared for different hardware platforms i.e. GPUs and FPGAs. We show that, even though GPU and FPGA have similar memory hierarchies, *the best implementation for an FPGA is obtained from OpenCL code that is different from the one leading to the best GPU implementation.* This is because the final selection of the *k* nearest neighbors is difficult to parallelize with the "doall" strategy implied by OpenCL, but it can still be efficiently pipelined on an FPGA. Furthermore, GPUs have higher DRAM access bandwidth as compared to an FPGA.

The FPGA implementation of our OpenCL code has been obtained by utilizing the SDAccel tool chain from Xilinx, including tools from the Vivado® Design Suite [6, 9]. The algorithm has been implemented on a Virtex-7 FPGA. The GPUs considered for comparison are the GeForce GTX 960 and the Quadro K4200, both by NVIDIA.

The main contributions of this paper are:

- An investigation of the issues encountered when implementing and optimizing an OpenCL code onto a Xilinx FPGA device.
- A performance comparison of the FPGA and GPU implementation in terms of time, energy and power.

The rest of the paper is organized as follows. Section II presents a brief overview of our algorithm and of the OpenCL programming model. A summary of the related work is presented in section III. Section IV describes in detail our adopted methodology. Section V presents the results and the work is concluded in section VI.

II. Overview

This section of the paper presents a brief overview of the kNN algorithm. An overview of OpenCL with a brief description of its platform, execution and memory model is also presented here.

A. kNN Algortihm

Given a set *S* of *n* reference (training) data points in a d-dimensional space and a query point *q*, the k-nearest neighbor algorithm returns the *k* points in *S* that are closest to point *q*. This is illustrated for k = 3 and n = 20 in Fig. 2. The circle represents the query point while the diamonds represent the reference data points.

The algorithm consists of the following main steps:

1- Compute *n* distances between the query point *q* and the *n* reference points of the set *S*. The distance in our case is the squared Euclidean distance, i.e. for two bi-dimensional points (x_1,y_1) and (x_2,y_2):

$$d = (x_1 - x_2)^2 + (y_1 - y_2)^2 \qquad (1)$$

2- Sort the *n* distances while preserving their original indices (as specified in *S*).

3- The *k* nearest neighbors would be the *k* points from the set *S* corresponding to the *k* lowest distances of the sorted distance array.

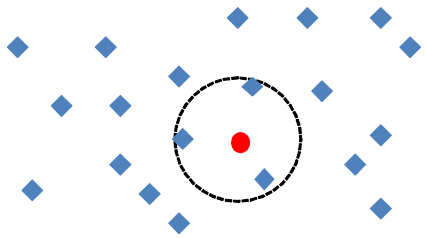


Fig. 2. Illustration of the kNN search algorithm with $k = 3$

B. Overview of OpenCL

OpenCL is an open, industry standard portable framework for writing parallel programs to be executed on heterogeneous platforms consisting of central processing units (CPUs), GPUs, digital signal processors (DSPs) and FPGAs [10]. OpenCL code can be run on a variety of supporting devices by making minimal changes to the host code, hence making it portable. The standard is derived from ISO C99 with additions to support both task-parallel and data-parallel programming models.

At the heart of the OpenCL platform model is the host, which is typically a CPU used to setup the environment for the OpenCL program to run on one or more devices. A device in OpenCL terminology is any hardware platform used to accelerate the compute intensive portions of the application. A piece of code running on the device is called a kernel. The OpenCL device consists of compute units (CU) each further divided into processing elements (PE) as shown in Fig. 3.

Several concurrent executions of the kernel body, called work-items, are grouped into work groups, which can be executed in parallel by multiple processing elements. The memory is broadly divided into host (i.e. CPU) memory and device (i.e. GPU or FPGA) memory. The device memory is also explicitly split into private memory (specific to each work-item), local memory (shared by all the work-items in a work group) and a global/constant memory shared by all the work groups. Global memory offers the slowest access but has the largest capacity while private memory is the smallest but the fastest among all. The memory model is also depicted in Fig. 3.

The main difference between OpenCL code executed on a CPU/GPU and an FPGA lies in the way the code is compiled. For CPU/GPU, the code is compiled in a just-in-time manner to exploit the fixed computing architectures of the devices. The intrinsic flexibility of the FPGA architecture on the other hand allows the designer to explore several kernel optimizations and CU combinations. The main caveat is that the generation of these highly optimized compute architectures takes longer than what a just-in-time compilation allows. The OpenCL standard addresses this issue by allowing for an

offline compilation of OpenCL code to be implemented on FPGA devices [6].

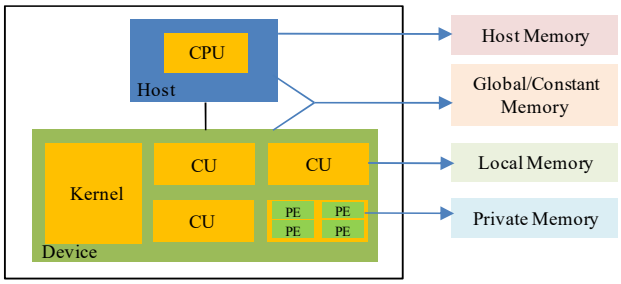


Fig. 3. OpenCL Platform and Memory Model

III. RELATED WORK

The prospect of using an FPGA as an accelerator in modern HPC systems has already been emphasized. This portion of the paper highlights some related work already done in this regard. It also describes some previous research work done in accelerating the kNN algorithm.

A performance comparison of a complex computer vision algorithm used for linear structure detection implemented over a GPU and an FPGA has been presented in [5]. The implementation platforms considered are an AMD Radeon HD6870 GPU and a Xilinx Spartan6 LX150 FPGA. The results show that the FPGA implementation performs better both in terms of power and speed as compared to the GPU. Unlike our case, where we use HLS to automatically generate RTL from an OpenCL code, the authors of this work have written VHDL code manually to implement the algorithm onto the FPGA. The complexity of writing code at RTL is obviously much higher than doing it for the GPU via OpenCL, because: (1) the number of lines of code is much higher at RTL, (2) verification and debugging are much slower, and (3) each RTL model implements one micro-architecture, while a single OpenCL model can generate several micro-architectures by providing different directives to the HLS tool e.g. pipeline or unroll a loop, partition a memory etc.

An accelerated kNN algorithm implemented on an FPGA-based heterogeneous computing system was presented in [11]. Altera's OpenCL compiler was used for compiling the OpenCL code onto the FPGA. The algorithm was implemented onto an Intel Core i7-3770 CPU, an AMD Radeon HD7950 GPU and an Altera's StratixIV 4SGX530 FPGA. The FPGA beats both GPU and CPU in terms of power and energy-per-computation consumption, but the GPU performs better than the FPGA in terms of execution time, most likely because of the higher DRAM access bandwidth of modern GPUs.

A detailed survey on how to parallelize the nearest neighbor algorithm was presented in [12] where, the author advocates both the opportunity and the need to parallelize such algorithms. A GPU based acceleration of a brute force kNN algorithm using CUDA and the CUBLAS library was presented in [7, 13]. It obviously showed a huge speed up with respect to a highly optimized C++ library implementation.

The use of FPGAs for acceleration is hence a widely accepted proposition. This is shown, e.g. by the decision by Baidu to accelerate its deep learning models for image search by using FPGAs [3]. Similarly, Microsoft, after years of research to accelerate its Bing search engine, is now also looking into how to accelerate deep learning models through FPGAs [4]. Considering the market demand, major FPGA manufacturers Altera and Xilinx have also recently introduced tools to program their respective FPGAs directly through OpenCL [6, 14]. One of our main objectives in this project is to explore how various OpenCL programming constructs are handled by the HLS tool and how can we extract maximum performance from the FPGA device.

IV. TEST CASE IMPLEMENTATIONS

In this paper, we compare two different OpenCL implementations of the kNN algorithm, and we show that the OpenCL code that leads to the best implementation on an FPGA is very different from the one that leads to the best GPU results.

A. Implementation I

The first implementation is a direct implementation of the most easily parallelizable part of the kNN algorithm, namely the distance calculation task, while both sorting and nearest neighbors identification are performed by the host. The implementation uses global memory only, and thus it is mainly a measure of global memory access bandwidth. The distance calculation for each point in the reference data set is completely independent of the other points making the algorithm extremely parallelizable. This implementation is illustrated in implementation I. It makes sense as an "acceleration" of kNN only if the dimensionality d of each point is high, and hence distance computation (which is $O(d*n)$) dominates over finding the k smallest distances (which is $O(k*n)$).

Implementation I. Distance calculation on device & neighbors on host

Input: A query point q and S, a set of reference points;
Output: Indices of the k reference points with the smallest distance from q;
Begin
On device:
1: **for** each reference point $s \in S$ **do**
2: compute all distances between q and all points $s \in S$;
3: **end for**
On host:
4: **for** $i = 0$ to $k-1$ **do**
5: print the index in S of the i-th smallest element of the distance vector;
6: **end for**
End

B. Implementation II

This implementation uses two separate kernels, and streams data between them. The first kernel is used to calculate the distances as in the previous case. The second kernel finds the k smallest distances and returns their indices at the end of its execution.

This implementation is meant to utilize a streaming memory optimization technique offered by SDAccel, which automatically maps global arrays, used merely for inter-kernel

communication, to the on-chip block RAMs. The pseudo code is given in implementation II.

Implementation II. kNN on device using multiple kernels

Input: A query point *q* and *S*, a set of reference points;
Output: *k* smallest distances with their respective indices per work-group;
Begin
On device:
1: declare global distance array "dist" for inter-kernel communication;
Kernel1: distance calculation
2: **for** each reference point *s* € *S* **do**
3: declare local arrays for each point *s* € *S*;
4: copy each point *s* € *S* into the local memory;
5: compute the distances between *q* and points *s* € *S* and save in "dist";
6: **end for**
Kernel2: find k smallest distances
7: **for** *i = 0* to *k-1* **do**
8 print the index in *S* of the *i-th* smallest element of the distance vector;
9: **end for**
End

V. RESULTS

We performed a comparison of our implementations on GPUs as well as an FPGA. The code has been optimized for the FPGA by using a variety of optimization options offered by the HLS tool from Xilinx, e.g. loop unrolling and pipelining, and we report the best results for each implementation on each platform. The experimental setup along with the results from our experiments is presented here.

A. Experimental setup

The experimental setup consists of three target devices shown in Table. I. The first device is an NVIDIA GeForce GTX960 GPU with 1024 cores and a maximum operating frequency of 1178MHz. The device has about 2GB GDDR5 of global memory, with 112GB/s of memory bandwidth. It is accessible from the host through a PCIe 3.0 interface with 16 lanes. The second device is an NVIDIA Quadro K4200 GPU with 1344 CUDA cores and a maximum clock frequency of 784MHz. The device has about 4GB of GDDR5 global memory, with 172.8GB/s of memory bandwidth. It is accessible from the host through a PCIe Gen2 interface with 16 lanes. The third device is an Alpha data ADM-PCIE-7V3 FPGA board with a Virtex-7 690t. The global memory consists of two DDR3 memories with 21.3GB/s of bandwidth. The host can access it through a PCIe Gen3 interface with 8 lanes.

TABLE I. TARGET PLATFORM COMPARISON

Device	Global memory size	Global Memory Bandwidth (GB/s)	Bus interface
GTX 960	2GB GDDR5	112.0	PCIe 3.0 x16
K4200	4GB GDDR5	172.8	PCIe 2.0 x16
FPGA	Two 8GB SODIMMs	21.3	PCIe 3.0 x8

The dataset used for experimenting with our kNN algorithm is from [15]. It contains locations of various hurricanes and is used by our algorithm to specify the *k* nearest hurricanes in the vicinity of a given query point. *k* is usually small in comparison to the number of points in the reference data set and we have fixed it to 5 in our experiments. The number of reference data points used is about 300,000.

B. Performance Analysis

The parallel architecture of the FPGA has been exploited by exposing parallelism in the kernels through several HLS optimization directives offered by SDAccel. The loop unroll attribute in SDAccel could be used to expose concurrency to the compiler by either fully or partially unrolling the loops in OpenCL kernels. However, fully unrolling loop iterations that access global memory, as in our case, does not ensure the best throughput, since only a few global memory ports are available. So we unroll only enough to match the available maximum number of global memory access ports. Throughput can be further improved by using the loop pipeline attribute, which can pipeline any explicit loop in the kernel as well as the work-item loop within a work group, and better match the limited number of memory ports.

The work group size in the OpenCL standard can be specified by the `(reqd_work_group_size)` attribute which indicates the size of the problem space that can be handled by a single invocation of the kernel compute unit. This attribute is highly recommended in the case of FPGA implementation, because it allows performance optimization during the custom logic generation for the kernel, by informing the synthesis tool about the iteration count of the loop over work items.

Several memory access optimizations are also offered by SDAccel which are critical to performance enhancements on an FPGA. For instance, 2-element vector data types improve the memory access throughput, as compared to using C structs, when reading in two-dimensional data points. One of the optimizations offered by SDAccel, namely "on-chip global memories", was exploited in implementation II to achieve a very significant speed up in execution. This optimization utilizes the block RAMs in the FPGA to create memory buffers that are visible only to the kernels accessing them, while inter-kernel buffers in "standard" OpenCL are allocated in the external, slower, DRAM.

The performance analysis for implementation I is presented in Table. II. The resource utilization in case of FPGA implementation is also shown. The frequency reported by Vivado HLS is 240MHz. The sorting in this implementation is done on the host. Hence, the sorting time in all the cases is also included as a part of the total execution time of the kNN algorithm. The devices in this implementation are used only to calculate all the distances between the query point and all the reference data points. This process is fully parallelizable with no loop dependencies. Work-item pipelining has been used here for FPGA implementation and the data is read in bursts from the global memory. Both the GPUs perform faster than the FPGA due to their higher DRAM access bandwidth. FPGA implementation however out-performs both the GPUs in terms of power and energy consumption. The power analysis for the FPGA implementation was done using the power estimation capabilities of Vivado. The GPU power on the other hand was estimated based on the datasheet, which from our earlier

experiments was very close to the one reported by GPU profiler tools e.g. GPU-Z.

TABLE II. PERFORMANCE ANALYSIS OF IMPLEMENTATION I

Parameters/Devices		FPGA	GTX 960	K4200
Device time		1.24ms	0.04ms	0.05ms
Sort time (Host)		4ms	3ms	3ms
Total execution time		5.24ms	3.04ms	3.05ms
Power (Device)		0.422W	120W	108W
Energy (Device)		0.523mJ	4.4mJ	5.6mJ
Resource Utilization	BRAMs	0	N/A	N/A
	DSPs	12 (0.33%)		
	FFs	3109 (0.36%)		
	LUTs	2006 (0.46%)		

The performance analysis for implementation II is given in Table. III. This implementation also has a clock frequency of 240MHz. It exploits the "on-chip global memories" option, offered by SDAccel for streaming data between kernels. A global memory buffer "dist" is used for inter-kernel communication which gets mapped to the on-chip Block RAMs and is visible only to the kernels that uses it. This explains the increased power consumption in comparison to the other case, where the BRAMs were kept powered down.

TABLE III. PEFORMANCE ANALYSIS OF IMPLEMENTATION II

Parameters/Devices		FPGA	GTX 960	K4200
Total execution time		1.23ms	0.93s	3.11s
Power		3.136W	120W	108W
Energy		0.0039J	111.6J	335.88J
Resource Utilization	BRAMs	512 (34.83%)	N/A	N/A
	DSPs	12 (0.33%)		
	FFs	23892(2.78%)		
	LUTs	11838 (2.76%)		

This implementation is considerably faster on the FPGA than both the GPUs, yet it still consumes both less power and less total energy in comparison. This speed up occurs at the cost of about 7x increase in the power consumption as compared to the FPGA implementation of implementation I. The best case GPU implementation (GTX960) in this case is about 756x slower than the FPGA implementation, since the multiple kernels execute sequentially on the GPU and share only global memory.

VI. CONCLUSIONS AND FUTURE WORK

This paper explores both kernel implementation changes and a variety of HLS directives to optimize the synthesis of an OpenCL application to be implemented on an FPGA platform. Two fairly different implementations of the kNN classification algorithm have been considered as our test cases. The FPGA is found to offer a better power/energy performance as compared to the GPU in all the algorithm implementations. By carefully analyzing the algorithm characteristics, we managed to find an OpenCL implementation of kNN that also results in better overall execution time on an FPGA than on a GPU, and is thus pareto-optimal with respect to GPU implementations with respect to performance, power and energy. It exploits on-chip global memory implementation and data streaming options that are more readily and more frequently available on an FPGA than on a GPU. Future work includes using our findings to enhance HLS tools to improve the level of automation starting from a non hardware specific OpenCL model.

ACKNOWLEDGMENT

The authors would like to extend their gratitude to Xilinx, Inc. for their support while carrying out this work. This work is also supported in part by the European Commission through the ECOSCALE project (H2020-ICT-671632).

REFERENCES

[1] Mavroidis, I., Papaefstathiou, I., Lavagno, L., Nikolopoulos, D. S., Koch, D., Goodacre, J., ... & Palomino, M. (2016, March). ECOSCALE: Reconfigurable computing and runtime system for future exascale systems. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 696-701). IEEE.

[2] Ovtcharov, K., Ruwase, O., Kim, J. Y., Fowers, J., Strauss, K., & Chung, E. S. (2015). Accelerating deep convolutional neural networks using specialized hardware. *Microsoft Research Whitepaper*, *2*.

[3] Ouyang, J., Lin, S., Qi, W., Wang, Y., Yu, B., & Jiang, S. (2014, August). Sda: Software-defined accelerator for largescale dnn systems. In *Hot Chips*(Vol. 26).

[4] http://www.nextplatform.com/2015/08/27/microsoft-extends-fpga-reach-from-bing-to-deep-learning/ [Accessed: 26 April 2016].

[5] Struyf, L., De Beugher, S., Van Uytsel, D. H., Kanters, F., & Goedemé, T. (2014, January). The battle of the giants: a case study of GPU vs FPGA optimisation for real-time image processing. In *Proceedings PECCS 2014*(Vol. 1, pp. 112-119). VISIGRAPP.

[6] User Guide, "SDAccel Development Environment User Guide v2015.1", Xilinx, 2015.

[7] Garcia, V., Debreuve, E., Nielsen, F., & Barlaud, M. (2010, September). K-nearest neighbor search: Fast GPU-based implementations and application to high-dimensional feature matching. In *2010 IEEE International Conference on Image Processing* (pp. 3757-3760). IEEE, http://dx.doi.org/10.1109/ICIP.2010.5654017.

[8] Che, S., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J. W., Lee, S. H., & Skadron, K. (2009, October). Rodinia: A benchmark suite for heterogeneous computing. In *Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on* (pp. 44-54). IEEE, http://dx.doi.org/10.1109/IISWC.2009.5306797.

[9] User Guide, "Vivado Design Suite User Guide High-Level Synthesis v2015.1", Xilinx, 2015.

[10] Stone, J. E., Gohara, D., & Shi, G. (2010). OpenCL: A parallel programming standard for heterogeneous computing systems. *Computing in science & engineering*, *12*(1-3), 66-73, http://dx.doi.org/10.1109/MCSE.2010.69.

[11] Pu, Y., Peng, J., Huang, L., & Chen, J. (2015, May). An efficient KNN algorithm implemented on FPGA based heterogeneous computing system using OpenCL. In *Field-Programmable Custom Computing Machines (FCCM), 2015 IEEE 23rd Annual International Symposium on* (pp. 167-170). IEEE, http://dx.doi.org/10.1109/FCCM.2015.7.

[12] Aydin, B. E. R. K. A. Y. (2014). Parallel algorithms on nearest neighbor search. *Survey paper, Georgia State University*.

[13] Garcia, V., Debreuve, E., & Barlaud, M. (2008, June). Fast k nearest neighbor search using GPU. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (pp. 1-6). IEEE, http://dx.doi.org/10.1109/CVPRW.2008.4563100.

[14] Singh, D. (2011). Implementing FPGA design with the OpenCL standard.*Altera whitepaper*.

[15] http://weather.unisys.com/hurricane/atlantic/2012/index.php [Accessed: 30 June 2016].

Education, Curricula & Research Methods

ECRM is a FedCSIS conference area aiming at interchange of information, ideas, new viewpoints and research undertakings related to university education and curricula as well as recommended methods of doing research in all computing disciplines, i.e. computer science, computer engineering, software engineering, information technology, and information systems. This area spans typical FedCSIS events (conferences, workshops, etc.) with rigorous paper submissions and review processes as well as panels, PhD and research consortia, summer schools, etc. Events that constitute ECRM are:

- IEES'16 - 1st International E-education Symposium—Education of the Future
- DS-RAIT'16—3rd Doctoral Symposium on Recent Advances in Information Technology

3rd Doctoral Symposium on Recent Advances in Information Technology

THE third international Doctoral Symposium on Recent Advances in Information Technology (DS-RAIT 2016) will be held as a satellite event of the Federated Conference on Computer Science and Information Systems (FedCSIS 2016) and Education, Curricula & Research Methods (ECRM 2016) conference.

The aim of this meeting is to provide a platform for exchange of ideas between early-stage researchers, in Computer Science, PhD students in particular. Furthermore, the symposium will provide all participants an opportunity to get feedback on their studies from experienced members of the IT research community invited to chair all DS-RAIT thematic sessions. Therefore, submission of research proposals with limited preliminary results is strongly encouraged.

Besides receiving specific advice for their contributions all participants will be invited to attend plenary lectures on conducting high-quality research studies, excellence in scientific writing and issues related to intellectual property in IT research. Authors of the two most outstanding submissions will have a possibility to present their papers in a form of short plenary lecture.

TOPICS

- Automatic Control and Robotics
- Bioinformatics
- Cloud, GPU and Parallel Computing
- Cognitive Science
- Computer Networks
- Computational Intelligence
- Cryptography
- Data Mining and Data Visualization
- Database Management Systems
- Expert Systems
- Image Processing and Computer Animation
- Information Theory
- Machine Learning
- Natural Language Processing
- Numerical Analysis
- Operating Systems
- Pattern Recognition
- Scientific Computing
- Software Engineering

EVENT CHAIRS

- **Kowalski, Piotr Andrzej,** Systems Research Institute, Polish Academy of Sciences; AGH University of Science and Technology, Poland
- **Lukasik, Szymon,** Systems Research Institute, Polish Academy of Sciences, AGH University of Science and Technology, Poland

PROGRAM COMMITTEE

- **Arabas, Jaroslaw,** Warsaw University of Technology, Poland
- **Atanassov, Krassimir T.,** Bulgarian Academy of Sciences, Bulgaria
- **Balazs, Krisztian,** Budapest University of Technology and Economics, Hungary
- **Bronselaer, Antoon,** Department of Telecommunications and Information at Ghent University, Belgium
- **Castrillon-Santana, Modesto,** University of Las Palmas de Gran Canaria, Spain
- **Charytanowicz, Malgorzata,** Catholic University of Lublin, Poland
- **Corpetti, Thomas,** University of Rennes, France
- **Courty, Nicolas,** University of Bretagne Sud, France
- **De Tré, Guy,** Faculty of Engineering and Architecture at Ghent University, Belgium
- **Fonseca, José Manuel**
- **Fournier-Viger, Philippe,** University of Moncton, Canada
- **Gil, David,** University of Alicante, Spain
- **Herrera Viedma, Enrique,** University of Granada, Spain
- **Hu, Bao-Gang,** Institute of Automation, Chinese Academy of Sciences, China
- **Koczy, Laszlo,** Szechenyi Istvan University, Hungary
- **Kokosinski, Zbigniew,** Cracow University of Technology, Poland
- **Krawiec, Krzysztof,** Poznan University of Technology, Poland
- **Kulczycki, Piotr,** Systems Research Institute, Polish Academy of Sciences, Poland
- **Kusy, Maciej,** Rzeszow University of Technology, Poland
- **Lilik, Ferenc,** Szechenyi Istvan University, Hungary
- **Lovassy, Rita,** Obuda University, Hungary
- **Malecki, Piotr,** Institute of Nuclear Physics PAN, Poland
- **Mesiar, Radko,** Slovak University of Technology, Slovakia
- **Mora, André Damas**
- **Noguera i Clofent, Carles,** Institute of Information Theory and Automation (UTIA), Academy of Sciences of the Czech Republic, Czech Republic
- **Pamin, Jerzy,** Institute for Computational Civil Engineering, Cracow University of Technology, Poland

- **Petrik, Milan,** Masaryk University, Czech Republic
- **Ribeiro, Rita A.**
- **Sachenko, Anatoly,** Ternopil State Economic University, Ukraine
- **Samotyj, Volodymyr,** Lviv Polytechnic National University, Ukraine
- **Szafran, Bartlomiej,** Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Poland
- **Tormasi, Alex,** Szechenyi Istvan University, Hungary
- **Wei, Wei,** School of Computer science and engineering, Xi'an University of Technology, China
- **Wysocki, Marian,** Rzeszow University of Technology, Poland
- **Yang, Yujiu,** Tsinghua University, China
- **Zadrozny, Slawomir,** Systems Research Institute, Poland
- **Zajac, Mieczyslaw,** Cracow University of Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 151–156
DOI: 10.15439/2016F574
ACSIS, Vol. 9. ISSN 2300-5963

Medical reporting in web-based applications designed to meet regulatory and industry standards

Michał Madera
Rzeszów University of Technology
al. Powstańców Warszawy 12,
35-959 Rzeszów
Poland
Email: michalmadera@gmail.com

Rafał Tomoń
SoftSystem Sp. z o.o.
ul. Leszka Czarnego 6a,
35-615 Rzeszów
Poland
Email: rtomon@softsystem.pl

Piotr Lorenc
SoftSystem Sp. z o.o.
ul. Leszka Czarnego 6a,
35-615 Rzeszów
Poland
Email: plorenc@softsystem.pl

***Abstract*—Web based applications penetrate into every software domain. Even those reserved only to desktop programs are becoming available through web browsers now. This has brought real technical challenges to software developers. Medical programs are not different. In this paper we are proposing new approach to text processing for web browser based medical applications. We are focusing on entering text of medical interpretation which is very important and sensitive aspect of medical report creation. A number of products were reviewed to justify the need for research in this area. The developed approach integrates report assembling, presentation and diagnosis text processing in accordance with medical data safety regulations. We prove that proposed solution based on HTML5 Canvas can be applied to development of the most demanding pathology reporting applications.**

I. Introduction

DEVELOPMENT of web based applications for reporting in medicine brings variety of challenges. The need for online access to software is unquestionable today. Initiated by the American Recovery and Reinvestment Act (ARRA) by 2009 pertained to areas like patient electronic health record [1] or management reporting. In spite of problems [2] online access to medical software tends to cover all areas nowadays. We will focus on challenges with porting medical reporting systems to web based applications. For the use of this paper we generalize medical report document structure [3]. The reports we are considering here can be divided into the following parts:

1. The Patient Information part containing patient demographics and clinical information (medical record number, attending doctor, etc.),
2. The Diagnostic Tests part which refers to laboratory testing results (including historical results) and observations,
3. The Medical Diagnosis part which is medical interpretation text entered by health care professionals (pathologist, radiologist, medical laboratory scientist, etc.) that we will refer to as diagnostician,
4. Report Header and Footer sections that usually contains medical institution information.

This work was supported by SoftSystem Sp. z o.o.

General template for medical report and its parts is presented in Fig. 1.

There is a wide context of medical reporting and problems we faced working with such systems. In this paper we

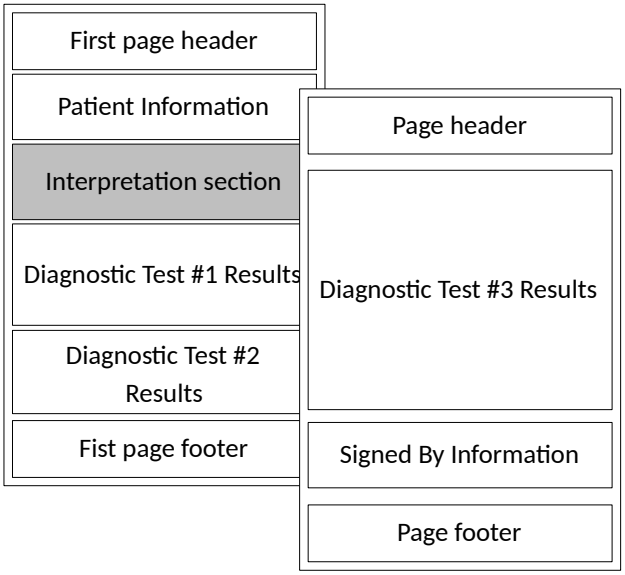


Fig 1. Medical report document structure

decided to focus on issues which are causing difficulties, especially in the web based environment.

Among the issues selected by Paul N. Valenstain, MD [4], the following are significant ones for which a solution will be presented in this paper:

1. When diagnostic tests are completed and report content is presented to a diagnostician, only the Interpretation sections can be modified (Interpretation can consist of multiple sections). All other sections of the document can not be modified (Diagnostic Tests, Patient Demographics and Report Header/Footer). Changing diagnostic or patient in-

formation would be considered as safety breach. We will refer to a problem of editable and not editable sections.

2. Complete Report layout has to be visible during interpretation entering and not changed when report is being printed or exported to PDF document. This is the concept of WYSIWYG (What You See Is What You Get).
3. Document presentation should not change across different web browsers, which is one of the most difficult issues to solve. Covering all notable web browsers and different release versions is one of the main requirement. Following browsers are taken into account: Chrome, FireFox, Opera, Safari, Internet Explorer and Microsoft Edge. Evaluation of components available on the market proved that research in this area is justified. Among others, following products were evaluated: DevExpress Html Editor [5], TxTextControl [6], TinyMCE [7]. All solutions based on ActiveX and Sliverlight technologies were not taken into consideration as these technologies are to be withdrawn.

II. Solution overview

The core of the research was to create software component providing the ability to view and edit medical reports in web browsers. It have to allow viewing the document in its final shape and editing in the same time, assuring the requirements listed in this paper's introduction are met.

Achievement was to develop web browser component based on HTML5 Canvas technology for text processing. Using Canvas for text processing is not a common approach with many obstacles that had to be overcome. We investigated possibility of using Canvas custom text drawing component in connection with innovative report merging mechanism. Medical data consistency and safety is guaranteed by presentation of not editable parts of the report as immutable images in conjunction with editable parts completely controlled by our canvas text processing editor. Immutable medical data images provides safety across all types of devices, while commonly used in web environment HTML rendering depends strongly on web browser engine and operating system. That may cause the content to be displayed differently on different devices, as depicted in chapter VI of this paper. For most of the content presented in web browsers this is not an issue. Unpredictable change to the formatting of medical data can lead to incorrect interpretation of the results. Using the same report merging mechanism to generate final PDF document and parts of the not editable report is the biggest gain of the presented solution. Reliability of presented information was the main goal and it was achieved.

III. Device independent document rendering

We defined the following components in our solution:

- *Reporting Engine* is a high-level text processing and reporting component for server-based application;
- *Web report* – final document divided into logical slices like header/footers, non-editable and editable slices;
- Shark Editor – text editor developed based on open source project named Carota.[8];
- *Shark JSON* – internal text editor format based on open-standard data format JSON defined by RFC 7159 [9];
- *Canvas* – element of HTML5 specification that allows dynamic, scriptable rendering of 2D shapes and bitmap images [10];
- *Font maps* – a set of properties described fonts like Times New Roman, Arial, Verdana, etc.

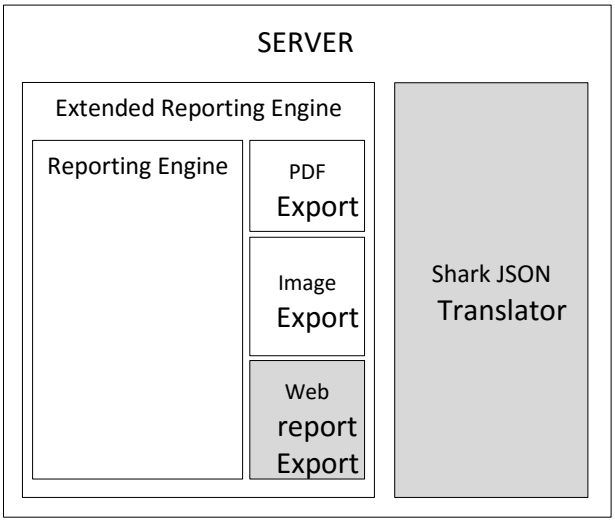


Fig 2. Server architecture

The architecture of the medical reporting system is a typical client-server architecture. System generates report on the server and returns it to the client using TCP/IP protocol. The client application, located on a web browser, presents the report to the user and handles all user inputs. Fig. 2 and Fig. 3 show graphical representation of the system architecture. The main element of server side application is a reporting engine which can provide fully merged medical report. The reporting engine was extended to provide the ability to export document in a new format – web report format. The advantage of this solution is fact that report engine is replaceable part of architecture. System may use any reporting engine available on market. Two requirements for the engine are: 1) ability to export document as image; 2) discreet representation of document in memory;

The client application was developed using HTML5 technology. Web report is passed to client in a JSON format, which is the most common data format used in browser/server communication. Text Editor Engine creates a virtual document with words, lines and sections, then measures it using provided font map, and then renders the document on the screen. Each element (word, letter, image, etc.) is drawn directly on the canvas that gives absolute control over what is presented on the screen and how it reacts to user changes.

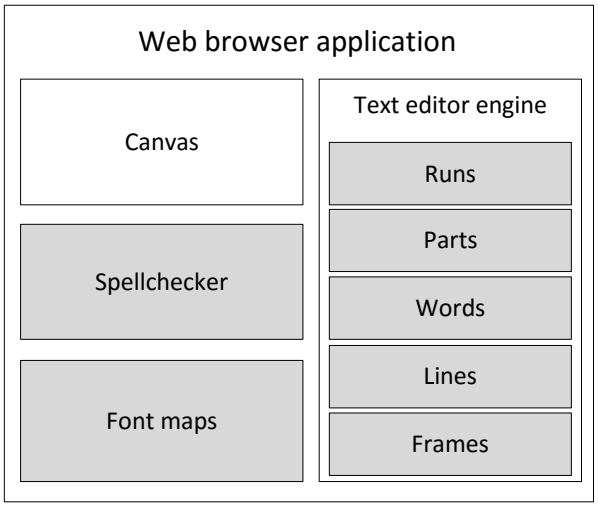


Fig 3. Web Client architecture

IV. Exporting the document

The web report is generated on server using the extended reporting engine. When the document is fully merged, the system is exporting it into bitmap image. Then using discreet data it is searching for logical slices: headers/footers, non-editable document content (diagnostic tests) and editable medical diagnoses.

The main attribute of logical slice is area boundary. This is a value calculated from the discreet representation of data. This attribute is used by server component to cut the image document exported earlier. Some of slices may occur on every page (like headers or footers) and to improve performance of exporting, the system cuts it only once. Cut slices are sorted in the order they occurred on the merged document. The generated web report contains information about whole document but it is divided into logical slices sorted chronologically.

Non-editable slices like headers/footers and diagnostic tests are exported as image fragments, but editable medical diagnoses are exported as RTF (Rich Text Format). System uses RTF, because RTF is a common format for text representation in medical software [11]. Diagnostic tests sections exported as image guarantee that the laboratory testing results will be modified neither by end user nor by malicious software.

The web report format may be used in text editors developed in other technologies like (X)HTML, WPF, Swing, WinForms. In our approach we use web report format to display it on canvas item and translate it to JSON format.

V. Canvas and document structure

Web Report in JSON format consists primarily of an array of JSON objects, called runs. They are parts of text with consistent formatting as well as some special elements (markups, images, etc). This format supports most commonly used properties, like font sizes and styles, colors and alignment. A sample text: “Plasma glucose is **about 12% greater** than...” is presented in JSON as follows:

```
[
   {   "text": "Plasma glucose is " },
   {   "text": "about 12% greater",
       "bold": true },
   {   "text": " than..." }
]
```

Elements in curly brackets are called *runs* and text from the sample is represented by three *runs*. Besides such simple objects, document structure also supports two complex types of *runs*. The first of them is also text based, but provides custom behavior on user input, e.g. could be modified only using drop-down or editor pop-up. It can be used to merge and maintain always up to date patient details, or provide some predefined values. This element contains extra attributes like *type* and *fielddata*. Below is an example of such item in JSON format.

```
{  "text": " Patient Last Name",
   "type": "TEXTFIELD",
   "fielddata": {
      "type": "MERGEDFIELD",
      "source": "PatientLastName" } }
```

The second type of *run* is not printed directly on the screen, but provides logical division of the document and ability to assign custom behavior. Used mostly to define read-only/editable items or distinct different parts of the report:

```
{  "text": {  "$": "sectionStart" },
   "code": "_SEC_GROSS_DESC" },
{  "text": "Interpretation text" },
{  "text": {  "$": "sectionEnd" } }
```

While the document loads, collection of runs is transformed into virtual structure of characters, words and lines which are stored in memory, to improve rendering performance. Each of these items have measured their own boundaries using a provided font map with ascent/descent values as well as width/height dimensions (Fig. 4). They are determined by the size of nested objects, so height of a word comes from the height of its letters, and the width is a sum of

letters widths. Line dimensions are determined by the nested words, documents by nested lines.

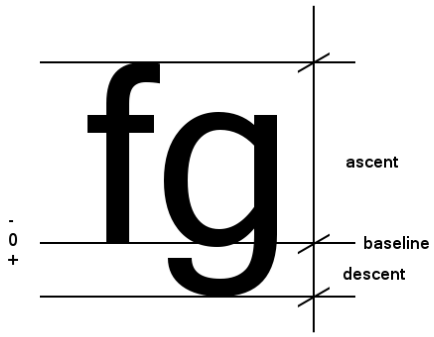


Fig 4. Font map to measure text size

Using font map not only speed up building of document structure, but also improve matching between canvas document and final PDF report.

JSON data also contains information about document layout, like paper size and margins, used to split content into words, lines and pages in the same manner, as in Reporting Engine. When the document structure is ready, visible part of the report (based on scroll position and window size) is drawn on the canvas, using its native methods like *fillText* or *drawImage* with measured sizes. Whole process starts from the document objects, that initiate drawing its lines and lines draws its words, etc. When the user is typing, entered text is first handled by hidden HTML text area component, which triggers document update – current word is immediately measured and the whole structure is updated.

Wrapping content and text between pages is divided into two steps, to maintain the best fit with final report. First, content is wrapped dynamically based only on current editor state (page size, line height, left space). This is fast, but can sometimes lead to unexpected results, such as splitting of non-editable, but consistent content, that should be moved as a whole to the next page. Fig. 5 demonstrate this situation.

When the user stops typing for a moment, a synchronization mechanism is triggered, as second step of this process. It requests the current pages split points from the PDF engine, compares them with current state and perform some minor additional adjustments, if needed. Fig. 6 demonstrate the result document after the synchronization. This mechanism guarantees that one of the postulates is met – "What You See Is What You Get".

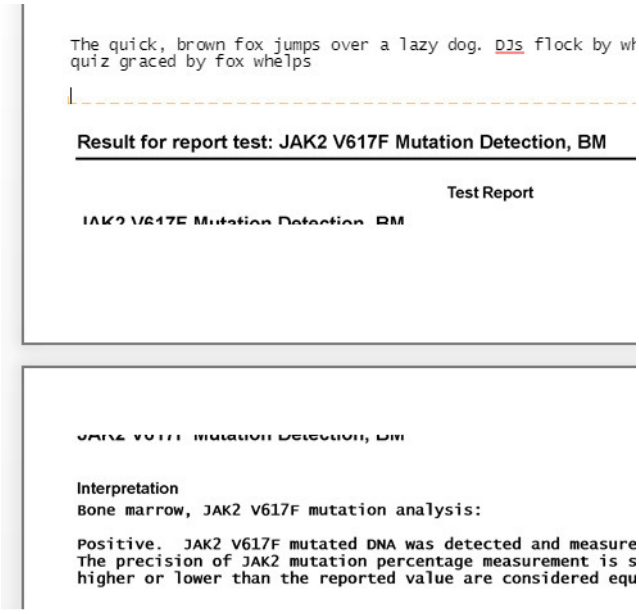


Fig 5. Page split while typing

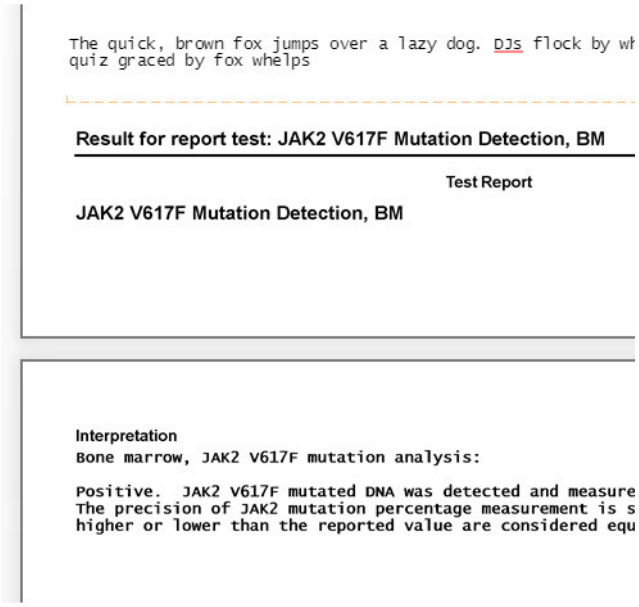


Fig 6. Page split resynchronized

VI. Evaluation

Evaluation of other components proved our solution to be superior. Taking leading product, Microsoft Word Online as an example, we prepared document with tabular data and opened in Internet Explorer browser Fig. 7. Difference in display when opened in Google Chrome is clearly visible in Fig. 8. Presented example demonstrates very dangerous change is data presentation. Due to different interpretation of HTML table size by web browser engine, numbers were con-

solidated in one line causing error prone situation. As presented in Fig. 9 the same document exported to PDF can change even more.

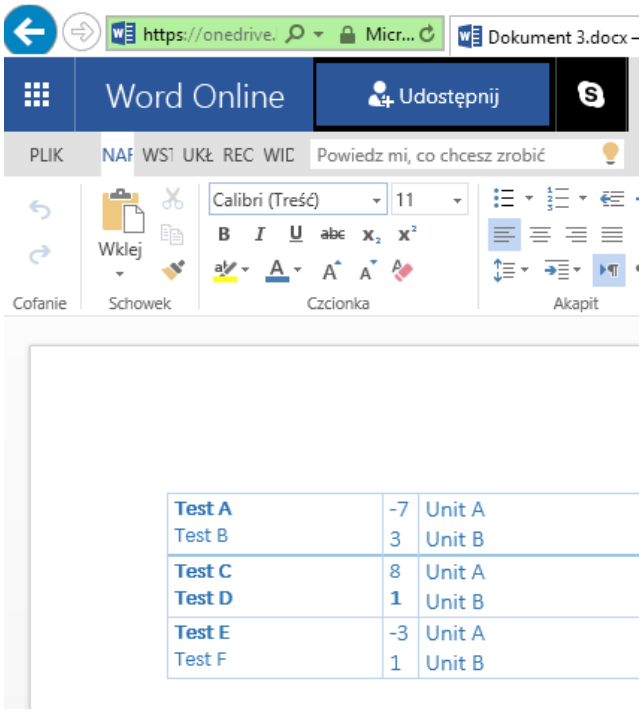


Fig 7. Document presented in Internet Explorer browser

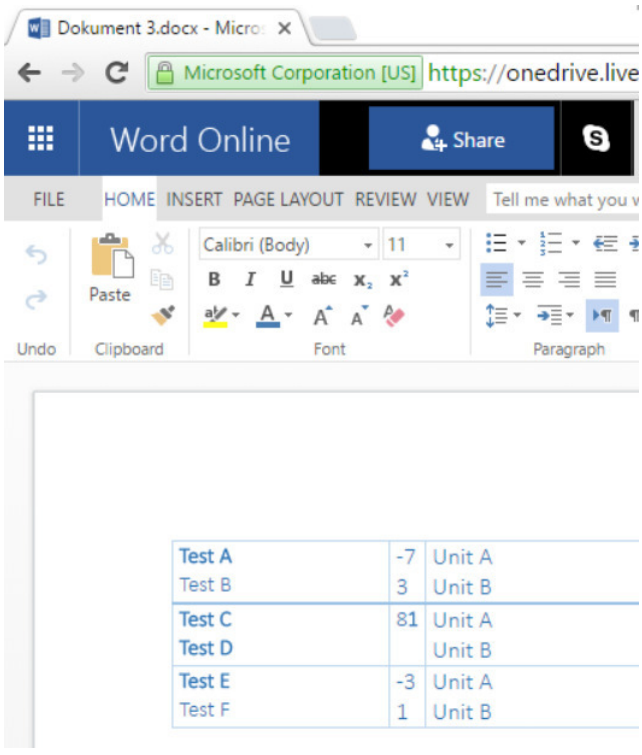


Fig 8. Document presented in Chrome browser

Such behavior is not acceptable for medical reporting. Described example would be classified as Risk to Health incident. Document processing we are presenting in our solution is designed to eliminate similar danger. Presenting of diagnostic data as images makes it platform independent.

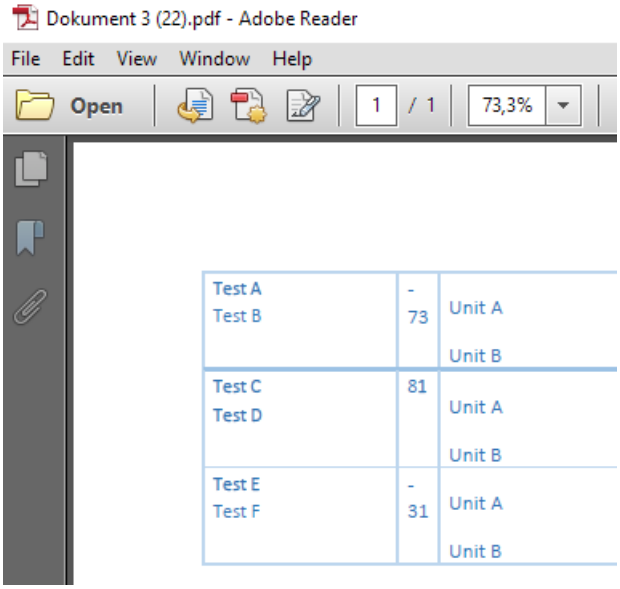


Fig 9. Document presented after exporting to PDF

VII. Results and Future Work

In this paper, a Report Editor Component for medical web applications was presented. We implemented the prototype and evaluated it in terms of useability. The component is fully functional with a satisfactory user experience. Comparison to Microsoft Word Online [12] and Google Documents [13] was done as these were considered to be a reference implementation of online text processing solutions. When editing large documents the response time has been evaluated as one of the most important aspect of usability. Significant performance improvement was achieved by the novel font mapping mechanism we developed. Support of medical dictionaries is added with “spell as you type” functionality. Initial evaluation of dictation was done but further work is necessary in this area. Still the main advantage of proposed solution is the safety of processed medical information. Design of the component provides extensible base for building sophisticated reporting software for medical industry.

One of the biggest challenges we have to face is table support. Handling tables, nested tables and table cells merging may be one of the most difficult task to complete. At this stage the component we designed can be provided to medical personnel for evaluation and suggestions. Feedback from pathologists would drive further work on this solution.

References

[1] T. Piliouras, A. Fortino, M. Andonov, and H. Huang, "Methodology to assist physicians in the selection of electronic health records software," in Applications and Technology Conference (LISAT), 2010 Long Island Systems, 2010, pp. 1–6.

[2] S. Ajami and T. Bagheri-Tadi, "Barriers for Adopting Electronic Health Records (EHRs) by Physicians," Acta Inform. Medica, vol. 21, no. 2, pp. 129–134, 2013.

[3] "Pathology Reports," National Cancer Institute. [Online]. Available: http://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet. [Accessed: 26-May-2016].

[4] P. N. Valenstein, "Formatting pathology reports: applying four design principles to improve communication and patient safety," Arch. Pathol. Lab. Med., vol. 132, no. 1, pp. 84–94, Jan. 2008.

[5] "Rich-Text and Html WYSIWYG Content Editing - ASP.NET MVC HTML Rich-Text Content Editor Demo | DevExpress." [Online]. Available: https://demos.devexpress.com/MVCxHTMLEditorDemos/Features/Features. [Accessed: 26-May-2016].

[6] "Text Control - .NET Reporting and Word Processing Components for Developers of Windows, Web and Mobile Applications (TX Text Control) | www.textcontrol.com." [Online]. Available: http://www.textcontrol.com/en_US/. [Accessed: 26-May-2016].

[7] "TinyMCE | The Most Advanced WYSIWYG HTML Editor." [Online]. Available: https://www.tinymce.com/. [Accessed: 26-May-2016].

[8] "danielearwicker/carota," GitHub. [Online]. Available: https://github.com/danielearwicker/carota. [Accessed: 26-May-2016].

[9] T. Bray, "The JavaScript Object Notation (JSON) Data Interchange Format." [Online]. Available: https://tools.ietf.org/html/rfc7159. [Accessed: 26-May-2016].

[10] "HTML Canvas 2D Context." [Online]. Available: https://www.w3.org/TR/2dcontext/. [Accessed: 26-May-2016].

[11] By Laura Bryan CMT BS Technology for the Medical Transcriptionist. Lippincott Williams & Wilkins Publishers, 2009.

[12] "Microsoft Word Online - Work together on Word documents." [Online]. Available: https://office.live.com/start/Word.aspx. [Accessed: 30-May-2016].

[13] "Dokumenty Google – twórz i edytuj bezpłatnie dokumenty online." [Online]. Available: https://www.google.pl/intl/pl/docs/about/. [Accessed: 30-May-2016].

4th International Conference on Innovative Network Systems and Applications

MODERN network systems encompass a wide range of solutions and technologies, including wireless and wired networks, network systems, services and applications. This results in numerous active research areas oriented towards various technical, scientific and social aspects of network systems and applications. The primary objective of Innovative Network Systems and Applications (iNetSApp) conference is to group network-related events and promote synergy between different fields of network-related research. To stimulate the cooperation between commercial research community and academia, the conference is co-organised by Research and Development Centre Orange Labs Poland and leading universities from Poland, Slovak Republic and United Arab Emirates.

The conference continues the experience of Frontiers in Network Applications and Network Systems (FINANS), International Conference on Wireless Sensor Networks (WSN), and International Symposium on Web Services (WSS). As in the previous years, not only research papers, but also papers summarising the development of innovative network systems and applications are welcome.

- EAIS'16—3rd Workshop on Emerging Aspects in Information Security
- SoFast-WS'16—5th International Symposium on Frontiers in Network Applications, Network Systems and Web Services
- WSN'16 - 5th International Conference on Wireless Sensor Networks
- main iNetSApp'16 track includes remaining topics, related to network systems and addressed not only to one of the tracks listed above

TOPICS

- Architecture, scalability and security of network systems,
- Web services – standards and applications,
- Service delivery platforms—architecture and applications,
- The applications of intelligent techniques in network systems,
- Innovative network applications,
- Network-based computing systems,
- Network-based data storage systems,
- Technical and social aspects of Open API and open data,
- Computer forensic and network security,
- Social, organizational and other aspects of information security,
- Network, cloud and data security,
- Misuse and intrusion detection,
- Traffic classification algorithms and techniques,
- Network protocols and standards,
- Network traffic engineering,
- Wireless communications,
- Control of networks,
- Internet of things,
- Sensor Circuits and Sensor Devices,
- Architectures, Protocols and Algorithms of Sensor Networks,
- Management, Energy and Control of Sensor Networks,
- Data Allocation and Information Processing in Sensor Networks,
- Resource Allocation, Services, QoS and Fault Tolerance in Sensor Networks
- Security and Monitoring of Sensor Networks,
- Software, Applications and Programming of Sensor Network,
- Performance, Simulation and Modeling of Sensor Network,
- Applications of Wireless Sensor Networks,
- Other aspects on network-related research.

STEERING COMMITTEE

- **Sebastian Grabowski, Research and Development Centre Orange Labs Poland,** Poland
- **Zakaria Maamar, Zayed University,** United Arab Emirates
- **Bohdan Macukow, Faculty of Mathematics and Information Science of the Warsaw University of Technology,** Poland
- **Juraj Miček, Department of Technical Cybernetics,** University of Žilina, Slovakia
- **Zbigniew Zielinski, Faculty of Cybernetics of Military University of Technology,** Poland

EVENT CHAIRS

- **Furtak, Janusz,** Military University of Technology, Poland
- **Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland
- **Hodoň, Michal,** University of Žilina, Slovakia

PROGRAM COMMITTEE

- **AbdAllah, Mohamed Mostafa,** Yanbu Industrial College, Saudi Arabia
- **Al-Anbuky, Adnan,** Auckland University of Technology, New Zealand

- **Baranov, Alexander,** Russian State University of Aviation Technology
- **Bataineh, Emad,** Zayed University
- **Ben-Othman, Jalel,** Université Paris 13
- **Chung, Danny Wen-Yaw,** Chung Yuan Christian University
- **Dabrowski, Andrzej,** Warsaw University of Technology, Poland
- **Fouchal, Hacene,** University of Reims Champagne-Ardenne, France
- **Fowler, Scott,** Linköping University
- **Frankowski, Jacek,** Orange Labs, Poland
- **Furnell, Steven,** Plymouth University, United Kingdom
- **Geiger, Gebhard,** Technical University of Munich, Faculty of Economics
- **Ghamri-Doudane, Yacine,** Université La Rochelle
- **Grabowski, Sebastian,** Research and Development Centre Orange Labs Poland, Poland
- **Gu, Yu,** National Institute of Informatics, Japan
- **Guedria, Lotfi,** Centre d'Excellence en Technologies de l'Information et de la Communication
- **Habbas, Zineb,** University of Lorraine
- **Haemmerli, Bernhard,** Hochschule für Technik+Architektur (HTA), Leiter Cisco Regional Academy, Switzerland
- **Howells, Gareth,** University of Kent
- **Husár, Peter,** Technische Universität Ilmenau, Germany
- **Jin, Jiong,** Swinburne University of Technology, Australia
- **Kaczmarski, Krzysztof,** Warsaw University of Technology, Poland
- **Kamoun, Faouzi,** Zayed University
- **Kiedrowicz, Maciej,** Military University of Technology, Poland
- **Kowalski, Andrzej,** Orange Labs, Poland
- **Ksentini, Adlen,** Université de Rennes
- **Laqua, Daniel,** Technische Universität Ilmenau, Germany
- **Legierski, Jarosław,** Orange Labs Poland, Poland
- **Macukow, Bohdan,** Warsaw University of Technology, Poland
- **Marir, Farhi,** Zayed University
- **Micek, Juraj,** University of Žilina, Slovakia
- **Milanová, Jana,** University of Žilina, Slovakia
- **Mokdad, Lynda,** Université Paris-Est, France
- **Monov, Vladimir V.,** Bulgarian Academy of Sciences, Bulgaria
- **Nowicki, Tadeusz,** Military University of Technology, Poland
- **Ouadoudi, Zytoune,** Université IbnTofail, Kenitra, Morocco
- **Scholz, Bernhard,** The University of Sydney, Australia
- **Ševčík, Peter,** University of Žilina, Slovakia
- **Shaaban, Eman,** Ain-Shams university, Egypt
- **Staub, Thomas,** Data Fusion Research Center (DFRC) AG, Switzerland
- **Stokłosa, Janusz,** Poznań University of Technology, Poland
- **Szmit, Maciej,** Orange Labs Poland, Poland
- **Xiao, Yang,** The University of Alabama, United States
- **Yang, Mee Loong,** Auckland University of Technology, New Zealand
- **Zieliński, Zbigniew,** Military University of Technology, Poland

5th International Symposium on 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

- 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
- **Nowicki, Tadeusz,** Military University of Technology, Poland
- **Richomme, Morgan,** Orange Labs, France
- **Wrona, Konrad,** NATO Consultation, Netherlands
- **Zieliński, Zbigniew,** Military University of Technology, Poland
- **Żorski, Witold,** Military University of Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 161–164
DOI: 10.15439/2016F503
ACSIS, Vol. 9. ISSN 2300-5963

QueuePredict – accurate prediction of queue length in public service offices on the basis of Open Urban Data APIs

Piotr Wawrzyniak
Research and Development Center
Orange Labs Poland
ul. Obrzeżna 7, 02-691 Warsaw, Poland
Email: piotr.wawrzyniak@orange.com

Jarosław Legierski
Research and Development Center
Orange Labs Poland
ul. Obrzeżna 7, 02-691 Warsaw, Poland
Email: jaroslaw.legierski@orange.com

Warsaw University of Technology, Faculty of Mathematics and Information Science,
ul. Koszykowa 75, 00-662 Warsaw, Poland

***Abstract* — This paper presents the methods to predict the number of people waiting in queues in Districts Offices of the City of Warsaw. On the basis of information from real-time queues length exposed as a part of Open City Data portal we can predict number of people in given time frame, which can be then used to further estimate predicted waiting time. These information are important and useful for number of people that visit district offices every day. The methods presented in the paper can be used to build a new value-added smart city services.**

I. Introduction

Queue management systems are popular tools in many organizations focused on mass customers service. When services are operated by limited number of support staff and at the same time we have to deal with a large, often unpredictably bursting number of visitors, we can expect the phenomenon of queues. Queue management systems (QMSs) can be useful in organizing the queues and very often shortening them. QMSs became popular solution used in crowd management all over the world, in Poland being especially popular in retail trade (e.g. in pharmacies), financial services (banks), electricity suppliers or in public offices (post offices, District Offices, Registry Offices etc.).

In this paper we will focus on the QMSs data provided by the City of Warsaw as a part of Open Urban Data platform of the City of Warsaw [1]. The remaining part of this paper is organized as follows:

- First, queue management systems present in the citizen servicing areas are described together with QMS API.
- Next, several applications that uses the API are described, with emphasis on the functionality and usefulness.
- Sections V and VI describe data collection process and prediction model assumptions respectively
- Finally prediction model evaluation and description of future works are provided.

II. Existing solutions

Citizen servicing offices of the City of Warsaw are served with multiple queue systems. When a citizen enters the office, he gets an appropriate ticket at a kiosk, and then waits until his number appears on the wallboard that manages the queue. Warsaw offices mainly uses QMS delivered by Swedish company Qmatic [2]. Installed solutions allows to transfer ticket between queues [3] when, for example, citizen must pay for the documents at the cash desk. Another available function is ticket reservation e.g. using web page. In this case, the person who booked the ticket have to enter the office five minutes before the scheduled time and confirm his/her presence using dedicated function at kiosk [4]. Unconfirmed tickets are automatically canceled and removed from the system.

Fig. 1. Queue system in Białołęka Ditrict Office in Warsaw (kiosk and wallboards)

It should be pointed out that current queue management systems used in Warsaw do not provide citizens with predicted queue length. Instead, only current queue

parameters [5], such as number of waiting people or expected waiting time are presented.

III. Queue Data Exposition

In addition to the queue displays in offices or web pages, that were mentioned in section II, the City of Warsaw also provides Application Programming Interfaces (APIs) to its' QMSs. This RESTlike APIs are part of Open Data portal [1], and cover QMSs serving 8 city offices. The list of the offices as well as number of distinct queues registered within each office is presented in Table I.

TABLE I
QUEUES EXPOSED IN API FORM IN WARSAW

No.	Office	Number of Queues
1	District Office- Białołęka	15
2	Registry Office Falęcka	5
3	District Office- Bielany	21
4	District Office- Ochota	27
5	District Office- Wola	24
6	District Office- Żoliborz	15
7	Registry Office Andersa	10
8	Passport Office Starynkiewicza	6

Open Urban Data platform used by the City of Warsaw was developed within MUNDO project (Apps4Warsaw) [6]. The platform collects data from QMSs with the frequency of 1 minute.

IV. Existing Applications

Number of application using QMSs API of the CoW were developed for past 2 years. The very first one was 'StaczKolejkowy' [8] that was developed within Business Intelligence Hackathon API (BIHAPI) contest in 2014 [7]. Another example use of the API is e-kolejka application that was competing in BIHAPI 2015. Both are mobile applications for Android - based smartphones and allows users to visualize queues on the screen and notify them when their ticket is being called. It should be emphasized that 'Staczkolejowy' was one of the winners of the 2014 BIHAPI contest.

Queue API was also popular API among participants of Apps4Warsaw contest. The applications concepts submitted to the competition and prototypes built in the development phase of this contest are listed in Table II.

TABLE II
APPLICATIONS PROPOSALS BASED ON QUEUE API DEVELOPED IN APPS4WARSAW CONTEST

No.	Project name	Status
1	Staczkolejkowy [8]	Prototype
2	Caffe Kolejka [9]	Prototype
3	Pan tu nie stał [10]	Concept
4	LessQStress[11]	Concept
5	ShortQ [12]	Concept

V. Data Preparation

In order to properly handle the task of prediction of the queue length, two datasets were collected between 1st of April 2016 and 7th of May 2016. For each of the datasets real-time queue data for a total number of 681 queues defined in 6 district offices (named *Q1, ..., Q6*) were collected using 1 minute polling interval. The data were then split into two groups, one of them, containing data from 1st of April till 30th of April, formed learning set, while the data collected in May were used to evaluate the accuracy of prediction model.

It was observed that queue length of 0 (zero) is reported very frequently by the API. Moreover, in about 5% of queries internal server errors were reported.

For each of the set only queue data for working hours were used in further analyses.

Overall our dataset contained of about 16 millions of valid individual entries of single queue length.

VI. Queue Prediction

As mentioned in section V, individual records of queue length are often equal to 0. Therefore we decided that instead of predicting such an individual state, our methods will estimate mean queue length in a given time frame. This means that raw datasets described in Section V had to be aggregated with the use of time-based windows. We decided to use two non-overlapping aggregation windows of length being respectively: 5 minutes and 1 hour. The idea of data aggregation is illustrated in Fig. 2.

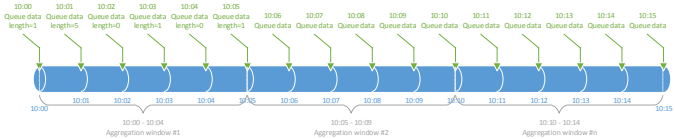


Fig. 2. Time-based windows used to aggregate raw queue data, window of 5 minutes used

For every window we calculated mean queue length, it means that in the example in Fig. 2 for window #1 the value assigned to this window will be 1.4.

In order to perform prediction of the queue length we decided to use Random Forests algorithm [13]. Input data for the prediction of queue length in time *t* were:

- aforementioned average queue length for *k* past aggregation windows, denoted *t-1, t-2, ..., t-k*,
- day of week (*d*) of aggregation window *t*
- time (hour and minutes) of the aggregation window *t*.

Due to methodology based on averaged values, we decided to use Random Forests in regression analysis task. For experimental verification of our method, we used three combinations of window size and *k* that are reflecting two use case scenarios: short-time predictions for people who are in a short distance to district offices and long-term prediction, for people traveling long distances to reach the office. Aggregation window length and number of historical observation for this use case scenarios are as follows:

- 5-minutes long window with *k=5* historical windows analyzed,
- 5-minutes long window with *k=12* historical windows analyzed,
- 60-minutes (1 hour) long window with *k=12* historical windows analyzed

VII. EXPERIMENT RESULTS

Proposed method prediction accuracy, defined as the proportion of predictions with error less than 0.1 person/timeframe, varies between 36.68% up to 88.16%, depending on district office. Mean prediction error is 0.714 person/timeframe in worst case, and 0.096 person/timeframe in the best case. It is important to note that distribution of error varies in time, as can be observed in the example at Fig. 3.

It can be also observed that almost all peaks observed in real data were predicted, while there are few false predictions, i.e. peaks that although were predicted by our algorithm, were never observed in real life data.

Our algorithm achieves the best results when shorter aggregation window of 5 minutes is used, despite the number of historical observations that are analyzed. The explanation for this fact is that usual peak in queue length lasts for around 30 minutes and therefore individual peaks are averaged over time window and produce more noisily data over which prediction is less accurate.

It should be also noticed that for the purpose of experiment presented in this paper we excluded data coming from hours were offices were closed (i.e. queue length was always 0). This means that prediction evaluation was performed in real use-case scenario.

Detailed comparison of obtained results for all three test cases are provided in tables III, IV and V respectively.

TABLE III
PREDICTION ACCURACY FOR AGGREGATION WINDOW OF 5 MINUTES AND *K*=5

Accuracy factor	O1	O2	O3	O4	O5	O6
Mean error (person/timeframe)	0,138	0,245	0,252	0,100	0,112	0,096
Standard deviation of error	0,353	0,879	0,917	0,355	0,302	0,377
Percentage of correct predictions	76,51%	67,48%	73,15%	88,16%	81,49%	86,69%

TABLE IV
PREDICTION ACCURACY FOR AGGREGATION WINDOW OF 5 MINUTES AND *K*=12

Accuracy factor	O1	O2	O3	O4	O5	O6
Mean error (person/timeframe)	0,156	0,278	0,269	0,121	0,123	0,117
Standard deviation of error	0,401	0,947	0,903	0,411	0,319	0,458
Percentage of correct predictions	77,92%	68,35%	70,21%	86,89%	81,46%	86,74%

TABLE V
PREDICTION ACCURACY FOR AGGREGATION WINDOW OF 60 MINUTES AND *K*=12

Accuracy factor	O1	O2	O3	O4	O5	O6
Mean error (person/timeframe)	0,392	0,602	0,714	0,277	0,250	0,180
Standard deviation of error	0,591	2,013	2,005	0,608	0,576	0,500
Percentage of correct predictions	42,04%	49,26%	36,68%	72,96%	48,23%	65,77%

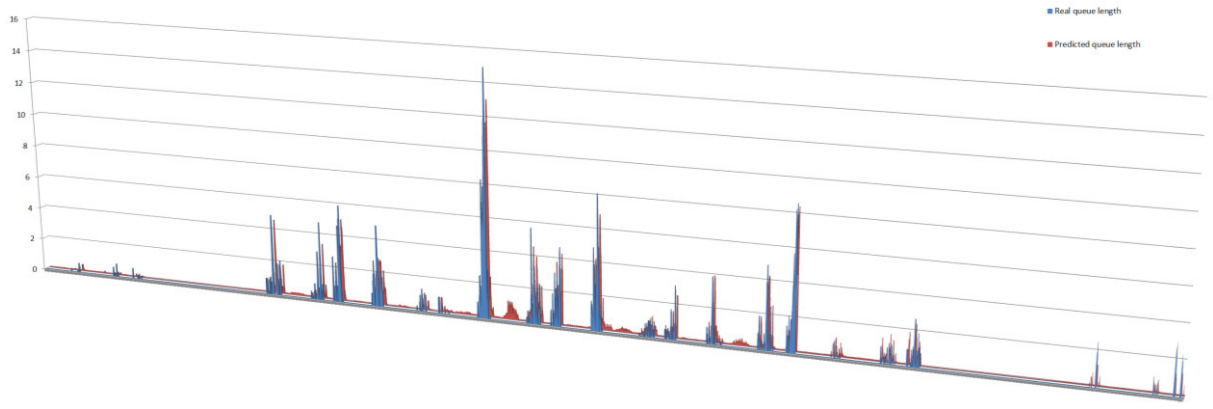


Fig. 3. An example comparison of real (blue) and predicted (red) queue length for 5 minutes long aggregation window for one of the district - offices.

VIII. Future Works

Currently, we put emphasis on development of accurate and computationally efficient prediction methods for queue data. Nevertheless proposed methods are designed to be easily transformed into QueuePredict API, which would be highly useful extension to already exposed urban APIs.

Moreover future works on prediction methods will focus on the use of additional data including but not limited to:

- weather data and weather forecasts,
- public transport data,
- long-term observation of queue usage,
- prediction of queue state in time steps far beyond near future (i.e. predicting queue length for several hours or days).

IX. Summary And Conclusion

This paper presented the evaluation of Random Forests regression model in the task of predicting the average length of queue in District Offices of the City of Warsaw. The queue data are collected with the use of Open Urban Data portal of the City of Warsaw and are collected in 6 district offices.

Prediction models were developed in three different scenarios representing different real life use cases. Prediction accuracy of the developed models is as high as 88.16 % with prediction error being as low as 0.096 person/timeframe when short time prediction is considered. Long term prediction facilitates significantly worse results with maximum prediction accuracy of 72.96 % and average prediction error of 0.180 person/timeframe.

X. Acknowledgment

This research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688380 'VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors'.

References

[1] City of Warsaw Open Urban Data platform, https://api.um.warszawa.pl/
[2] Qmatic company webpage, http://www.qmatic.com/
[3] Customer Experience Management, Qmatic white paper, Nov 2015, available online at http://lp.qmatic.com/the-customer-experience-orchestrated
[4] Białołęka District Office, visit registration webpage, available at http://bialoleka.waw.pl/strona-451-internetowa_rejestacja_wizyt_w_urzedzie.html
[5] Białołęka District Office, queue monitoring webpage, available at http://bialoleka.waw.pl/strona-450-sprawdz_kolejke.html
[6] Apps4Warsaw project page, available at http://www.apps4warsaw.org/
[7] Bihapi contest webpage, available at http://bihapi.pl/
[8] 'Staczkolejkowy', application description at Apps4Warsaw page, available online at https://konkurs.danepowarszawsku.pl/pl/projekt/stacz-kolejkowy
[9] 'Caffe kolejka', application description at Apps4Warsaw page, available online at https://konkurs.danepowarszawsku.pl/pl/projekt/caffe-kolejka
[10] 'Pan tu nie stał', application description at Apps4Warsaw page, available online at https://konkurs.danepowarszawsku.pl/pl/projekt/pan-tu-nie-stal
[11] 'lessQstress', application description at Apps4Warsaw page, available online at https://konkurs.danepowarszawsku.pl/pl/projekt/lessqstress
[12] 'shortQ', application description at Apps4Warsaw page, available online at https://konkurs.danepowarszawsku.pl/pl/projekt/shortq
[13] L. Breiman, "Random Forests," Machine Learning, 45, pp. 5-32, Kluwer Academic Publishers, 2001.

5th International Conference on Wireless Sensor Networks

TOPICS

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
- **Ševčík, 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
- **Fouchal, Hacene,** University of Reims Champagne-Ardenne, France
- **Furtak, Janusz,** Military University of Technology, Faculty of Cybernetics, Poland, Poland
- **Giusti, Alessandro,** CyRIC - Cyprus Research and Innovation Center, Cyprus
- **Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland
- **Gu, Yu,** National Institute of Informatics, Japan
- **Hudik, Martin,** University of Zilina
- **Husár, Peter,** Technische Universität Ilmenau, Germany
- **Jin, Jiong,** Swinburne University of Technology, Australia
- **Jurecka, Matus,** University of Žilina, Slovakia
- **Kafetzoglou, Stella,** National Technical University of Athens, Greece
- **Karastoyanov, Dimitar,** Bulgarian Academy of Sciences, Bulgaria
- **Karpiš, Ondrej,** University of Žilina, Slovakia
- **Kochláň, Michal,** University of Žilina, Slovakia
- **Laqua, Daniel,** Technische Universität Ilmenau, Germany
- **Milanová, Jana,** University of Žilina, Slovakia
- **Monov, Vladimir V.,** Bulgarian Academy of Sciences, Bulgaria
- **Ohashi, Masayoshi,** Advanced Telecomunications Research Institute International / Fukuoka University, Japan
- **Papaj, Jan,** Technical university of Košice, Slovakia
- **Ramadan, Rabie,** Cairo University, Egypt
- **Scholz, Bernhard,** The University of Sydney, Australia
- **Shaaban, Eman,** Ain-Shams university, Egypt
- **Shu, Lei,** Guangdong University of Petrochemical Technology, China
- **Smirnov, Alexander,** Linux-WSN, Linux Based Wireless Sensor Networks, Russia
- **Staub, Thomas,** Data Fusion Research Center (DFRC) AG, Switzerland
- **Teslyuk, Vasyl,** Lviv Polytechnic National University, Ukraine
- **Wang, Zhonglei,** Karlsruhe Intitute of Technology, Germany
- **Xiao, Yang,** The University of Alabama, United States

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 167–173
DOI: 10.15439/2016F580
ACSIS, Vol. 9. ISSN 2300-5963

Digital signing for short-message broadcasted traffic in BLE marketing channel

Jarogniew Rykowski
Poznań University of Economics and Business
Niepodległości 10
61-875 Poznań, Poland
email: rykowski@kti.ue.poznan.pl

Mateusz Nomańczuk
Billennium sp. z o.o.
Tagore'a 3
02-647 Warszawa, Poland

***Abstract*—As long as Bluetooth Low Energy (BLE) was mainly applied for broadcasting marketing information, the problem of trust of this transmission was treated as marginal. However, once the marketing channel was applied for such application as geolocation by means of BLE beacons, and e-payments, the problem of proper identification and authentication of the broadcasting device, as well as time&place of interaction, become very sharp. This problem cannot be solved by means of traditional mechanisms such as symmetric and asymmetric cryptography, due to several reasons. First, symmetric cryptography needs a redistribution of an encryption key, common for all the network nodes or at least known for the network central authentication point, and kept secret for the lifetime of the nodes. It is very problematic how to keep such multi-copied and long-lasting information secret. Second, the messages broadcasted in BLE marketing channel are restricted by length and format, making it practically impossible to use longer encryption keys widely assumed as safe. Third, BLE devices are usually very restricted according to memory amount and processing power, thus classical implementation of PKI encryption algorithms is very problematic. Fourth, there is no way to apply usual two-directional interaction to exchange some data to be encrypted, e.g., to proof directly the fact of interaction between two devices. And last but not least, time representation in small autonomous devices is quite weak, thus the hardware must be extended by some additional verification mechanisms and specialized hardware modules.**

In the paper we present a practical approach to an efficient representation of a testbed for trusted geolocation beacons broadcasting in the BLE marketing channel. The encryption is based on external co-processor and elliptic curves algorithms, which made it possible to apply shorten keys and use minimum resources of the beacon (memory, processor, energy). To prevent the attacks of "recording" type in man-in-the-middle mode (reusing the broadcasted information obtained in one place in the other place/time), the broadcasted messages include time stamps generated by attached RTC units. The idea may be applied for the other types of IoT and sensor networks to improve trust and verification of broadcasted messages.

This work was partially supported by the GOLIATH project jointly funded by the Poland NCBR and Luxembourg FNR Lead Agency agreement, under NCBR grant number POLLUX-II/1/2014 and Luxembourg National Research Fund grant number INTER/POLLUX/13/6335765.

This work was partially supported in scope of the project "Creating a platform to support management of Rego projects" co-financed by the Operational Programme Innovative Economy, action 1.4-4.1 - Support for research, development and implementation of the results, the grant agreement No. UDA-POIG.01.04.00-14-221 / 11 -00

I. Introduction

RECENTLY we observe a boom of Bluetooth devices and applications. Bluetooth (BT) communication is widely applied for short-distance networking, mainly for personal purposes, but also for some sensor networks [1]. With the introduction of version 4.1 (Bluetooth Low Energy BLE [2], recently also called Bluetooth Smart), apart traditional one-to-one transfer among two paired devices, it is also possible to generate and receive messages in the broadcast mode. Such a message, generated by one BLE node, may be read by any other BLE node, temporary present in the radio-range area, without the need of previous pairing or installation of some new software. Originally, BLE broadcast was applied for some marketing purposes. The idea was to generate periodically some short messages with advertisement of "local" products and services, to attract anybody in the closed neighborhood. This idea soon evolved towards a standard called BLE Marketing Channel. The standard concerns possible message formats, as well as physical parameters of the transmission – maximum power, message length and necessary parts (such as preamble and CRC), minimal gaps between transmissions (frequency of repetitions) etc. In a while BLE broadcast was applied to some other application areas, such as shopping and orientation in supermarkets. This idea in turn resulted in the introduction of beacons.A `beacon` is a small autonomous device, disconnected from the network, periodically (usually few times per second) broadcasting some information about itself, mainly unique identifier [3]. The identifier, in conjunction with an external database, may be used to deduce an exact geo-location of the device. As the transmission power of the device may be adjusted to current needs (at the installation time only, however), restricting the transmission distance

from centimeters to a few meters, it may be assumed that all the receivers in the radio-range share the same location as long as they are able to receive proper information from the beacon [4].

The above mentioned "receivers" are any BLE devices capable of reading broadcasted messages. To this goal, any modern smartphone or tablet would apply. In case of iOS devices, there is no need for an installation of any additional software – support for BLE traffic is included as basic functionality of the operating system. In case of Android, usually a dedicated application must be installed. For Windows, according to our best knowledge, there is no BLE support as for now.

Once the BLE technology is so popular, and both hardware and software is there, it is very probably that BLE broadcast will be used in many places and for many different purposes, such as sensor networks, automatic ticketing, e-payments [5], tracking etc. However, we have to enumerate not only the advantages, but also new problems provoked by possibly mass usage of this technology. As the main target or the BLE broadcast was addressed to the advertisement, such features as trust, privacy and security [6] were not initially taken into attention. As a consequence, the standard bypasses such basic functionality as digital signing of the broadcasting BLE nodes, encryption and decryption of the message content, verification of message consistency (apart standard CRC verification) etc.

The main goal of our work is to fill the gap. In the paper we discuss some possible ways of hardware and software extensions for geolocation beacons (and similar, any broadcasting node of a sensor network) to be used for any application which requires much more level of trust. To preliminary test the idea, we propose some ways for early-prototype development, based on linking BLE devices with Windows-operated PC, to use its full potential for finding the best algorithms for the encryption and decryption of the information to be broadcasted by the beacons.

The remainder of the paper is organized as follows. First, we briefly describe BLE marketing channel and data formats for geolocation beacons, followed by a discussion on the problem of a lack of efficient cryptography method for this sort of devices. Second, we try to formulate basic requirements for such method, to be applied mainly for digital signing of the broadcasted information. Then, we propose an environment for testing several solutions, with a prototype testbed based on AVR processors [7], BLE modules and some PC-based simulators of encryption modules. Finally, we provide some conclusions and show the directions of future work.

II. BLE MARKETING CHANNEL AND GEOLOCATION BEACONS

As mentioned already, a beacon uses so called BLE marketing channel to disseminate some information about itself, mainly unique identifier. The channel is characterized by some restrictions, especially introduced to minimize energy consumption and extend battery life. In particular, in any case message size cannot exceed 47 bytes, and certain time gap between messages must be preserved, making it possible to transmit only a few messages per second.

The above requirements substantially reduce the possibility to directly apply encryption for the broadcasted messages, for at least two reasons: limited data length, and limited possibility to mix both encrypted and non-encrypted data.

To better understand these restrictions, we must describe the data format that is used to disseminate messages in BLE marketing channel. Each message, of length 47 bytes, is composed of (Fig 1):

- a preamble, always equal to 0xAA,
- channel address (in turn always equal to 0x8E89BED6),
- data packet, composed of a header, additional address field, and unique device identifier (in case of the most popular iBeacon standard[8], for other standards such as AltBeacon [9], Radius Networks, Google beacons [10] etc., similar restrictions apply),
- strength of the radio signal (power level),
- standard Cyclic Redundancy Check (CRC) value to automatically detect and correct transmission errors.

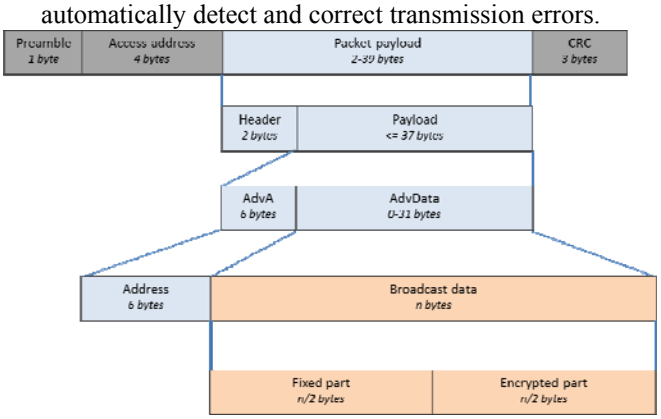


Fig.1. Single frame for BLE marketing channel, iBeacon standard

As it may be seen, only 31 of 47 bytes composing the message may vary, while the rest is reserved for the purposes of BLE marketing channel standard. This means that the maximum length of the encrypted part is limited to 248 bits. It is widely assumed that the encryption key must be shorter than the message to encrypt. Thus, in this case the encryption algorithm should use keys of minimum length, which are not treated as safe now. Note that parting the encrypted data into several messages is also very problematic, due to the fact the broadcasting takes place only few times per second, thus the reception of the whole message would be incredibly long. Note also that if we would like to transfer in the same message both encrypted and non-encrypted part (for the purposes of digital signing), the encrypted data length is limited to 20 bytes, thus 160 bits at the most.

Taking into account the above restriction, we decided to applied the newly proposed encryption algorithm based on

elliptic curves (ECC) [11]. This algorithm is effective for the substantially shorten keys (in comparison with classic algorithms such as RSA [12]), its 160-bit long key provides similar encryption power as 1024-bit RSA key. The problem is the ECA algorithm is hardly applicable to the limited hardware/software of the beacons. Even if successfully implemented with limited memory and CPU resources, it would consume substantial amount of energy for the computations. Thus, for the encryption a specialized chip should be applied, to (1) encrypt the data in parallel with the execution of a beacon program, and (2) to minimize energy consumption with the encryption process implemented at hardware level.

So far, only limited number of such specialized chips is available on the market; however, this restriction is expected to be relaxed. In parallel, several basic libraries have been proposed recently to apply ECC algorithms for the encryption, as DSA/RSA replacement. These libraries are available for popular programming languages such as Java and C/C++/C#. In the next section, we broaden the discussion on this topic, addressing also some implementation and organizational issues related to the usage of trusted (signed) information broadcasted by the beacons.

III. Theoretical Aspects of Encryption for Broadcasted Messages

In the classical application of a beacon to geo-location marking, each beacon broadcasts only its unique identifier, and no encryption is applied. Also, no time is to be represented and broadcasted – the broadcasted signal is the same for the lifetime of a beacon (usually from a few months to a few years). As such, the signal is not resistant to any man-in-the-middle attacks, such as capturing a message from a beacon at certain place and re-broadcast this signal using a fake transmitter at another place. Moreover, such attack seems to be trivial using popular devices such as smartphones with BLE units. It may be expected that sooner or later attacks of this type will take place, especially at the shopping centers and popular tourists points, for which the beacons are widely used as basic location markers.

It is clear that the beacons should be extended by some verification mechanism. And it is quite evident that the cryptography should be applied to this goal. The question is – which kind of cryptography and which algorithms?

As for the cryptography type, we deal with symmetric and asymmetric cryptography. The base for the first is to distribute some encryption keys, usually common for all the nodes in a local "network" (it's hard to say the broadcasting-only nodes create a full network, however, we will use the term "network" to point out the fact the nodes are communicating using networking mechanisms). The problem with such distribution and storage is to keep the keys secret. As the beacons are fixed and cannot be changed/updated, and they have no bi-directional network connection, it is very naïve to think the key would not be unfolded, and thus some fake devices would be added to the "network". Moreover, as most of nowadays smartphones are equipped with a BLE unit, it would be possible to run an application pretending to act as a beacon with minimum effort. For this reason, we should abandon the idea of using symmetric cryptography for beacon network.

The asymmetric cryptography seems to be much better candidate [13]. As the private key is possibly generated at the installation time inside the device and never accessible outside, there is no way to unfold it. And even the device is cracked by the hardware, it is possible only to falsify this device only, while the other devices are still safe and trusted [14]. The public key may be either propagated at the installation time to the operator, to be stored in an external database indexed by the device unique identifier [15], linked to IP address [16], or even periodically broadcasted by the beacon together with its identification data.

However, several problems arise while trying to implement asymmetric cryptography for BLE marketing channel, these are discussed below, with some proposals towards efficient implementation.

First, as already mentioned in the previous section, BLE marketing channel is very restricted according to total length and format of the broadcasted messages. Second, majority of information from each message is fixed due to the requirement of the BLE marketing channel standard. As a result, only small fraction of the message is variable, and only a small part may contain encrypted data. Thus, it is not possible to apply classical encryption algorithms such as RSA, for which the basic requirement is to encrypt the information longer than the encryption key. A reasonable key is 1024 bits long, while the minimum encrypted message length is something like 128 bytes (not counting fixed elements of each message such as headers and checksums), which is at least three times as much as the maximum length of a message in BLE marketing channel. One may say that a single message may be cloned to form a longer message, and such a long message may be divided into a few smaller messages, transmitted into pieces and finally relinked and extracted at the receiver side. However, due to energy savings, minimum period between succeeding transmissions is counted in hundreds of milliseconds. Thus, transmitting the whole long message would take a second or even more, assuming there will be no transmission errors. As the useful radio-distance for a beacon is sometimes counted in centimeters, one would have to stop and wait near each beacon to get the whole encrypted message. This is unrealistic, moreover, usually we do not know exact locations of the beacons, thus having no information where to stop and for how long.

So, if RSA (and similar solutions) cannot be directly applied, we have to take a look for its replacement. Elliptic Curve algorithms (ECC) seems to be a good candidate to this goal [11]. With their 160-bits long keys they offer

encryption power comparable with 1024-bits RSA [12]. Once a minimum length of an encrypted message is counted down to 20 bytes only, ECC encryption fits the maximum length of BLE message.

In addition, ECC algorithm is much less demanding according to the memory and CPU requirements in comparison with RSA [12]. Note that when we have to e.g., double the memory amount, we usually shorten the battery life by the same factor. Thus, for the autonomous beacons, for which the lifetime should be counted in years rather than months, it is extremely important to avoid complex computations and mass usage of operational memory. However, the problem is that typical processors applied for beacons and similar IoT/sensor network devices are very restricted. E.g., popular AVR family offers 8-bit processors with 2 kB of operational memory as the base for small battery-operated devices. Even if some AVR processors are 32-bits machines with hundreds of kilobytes of memory, it is still not reasonable to use such hardware for ECC computations. As stated by many researches, it takes typically several hundreds of milliseconds to perform a single encryption for a short message while using AVR-related hardware [7]. Our experiments also showed that efficient implementation of any asymmetric encryption algorithm for AVR processor, even if possible, takes too much time and leaves almost no place in the memory for the other code, needed at least to control the operation of BLE unit and message composition.

The solution is to apply a separated hardware module dedicated to encryption tasks. We tested such modules, but found their usage very difficult while applying for AVR-based nodes. Unfortunately, these modules are hardly re-programmable, and there is limited number of libraries for popular AVR software-design platforms. It looks like a lot of work should be done in order to propose a more practical solution. Anyway, we clearly see this is the right way, and we would like to test this way even if the external, fully programmable encryption modules are not ready yet. Our approach towards such a testbed in unfolded in the next section.

Finally, we have to discuss the strategy for direct beacon verification, and indirect verification of time/place of the interaction with the beacon. Usually, to prove the fact of such interaction, bi-directional transmission is used based on exchange of some encrypted and to-be-encrypted information [16, 17, 18]. For example, to prove the fact of being in the radio-range with a device, another device near-by composes a random message, encrypts it with public key of the device and sends to this device. The device, using its private key, decrypts the message, in turn encrypts it with the public key of another device and sends it back. Such double encrypted/decrypted message may be a proof of the cooperation of these two devices. If in addition encrypted time-stamps are exchanged rather than randomized messages, one is able to prove not only the fact of transmission, but also its location in time, proven by both parties.

As the above schema is based on bi-directional traffic, it cannot be applied for the broadcasted-only system. Instead, a different approach must be applied, aiming in using different broadcasted messages for subsequent transmissions. The messages must differ in such a way in reasonable time there will be no possibility to use the same device twice. As the geo-location message in the BLE marketing channel is sent few times per second, and the lifetime of the beacon is counted in years, there should be at least approximately 30 million of different combinations, which stands for a “long integer” value – not a problem even for a very small microcomputer. Moreover, this number may represent time (e.g., as number of seconds passed since the moment of installation), acting as a time stamp for proving the fact of interaction and never re-used. For better quality, Real Time Controller (RTC) module may be added, counting an reporting seconds (or even milliseconds, if needed) with an error not exceeding few seconds per year, which is acceptable for most of the applications of geo-location beacons.

In the next three sections we depict a testbed implementing a network of trusted geolocation beacons, composed of three levels:

- beacon-simulator level, based on Arduino hardware and software with dedicated BLE transmission module,
- central monitoring and encryption node, based either on extended Arduino controller with hardware encryption module, or a PC with encryption software,
- application level, aimed in using Android-controlled devices to test the behavior of the trusted beacons.

IV. Beacon simulators

In general, we were not able to directly apply any of market solution, i.e., commercially available beacons, to be used in the trusted mode. The reason is the beacons are closed solutions, with fixed behavior and with no possibility to re-program to suit our needs for the encryption and trust. Thus, since the very beginning we decided to use beacon simulators instead. A simulator is based on Arduino microcontroller with BLE transmission unit, real-time clock and encryption module attached (Fig. 2).

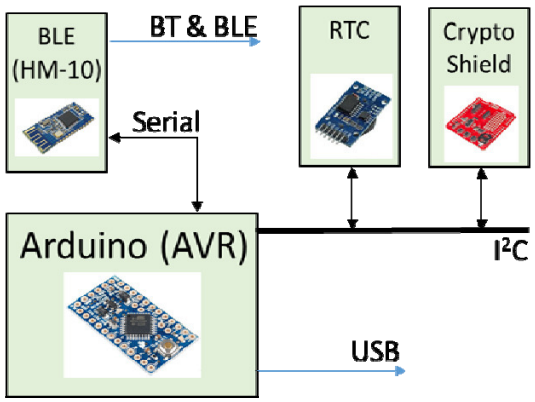


Fig.2. Beacon simulator

We have to point out here that AVR-based hardware, Arduino included, is very restricted according to efficient ways of code preparation and testing. Usually, the code is prepared using a PC-based compiler and then send to the target module by a communication link (typically Bluetooth or USB). There, the code is to be executed, usually with very limited possibilities of tracking and debugging. As a result, proper implementation of sophisticated tasks, such as asymmetric encryption protocols, takes a lot of time and programmers' efforts and requires high programming skills.

The operation mode of a beacon simulator was the following. First, RTC time was fetched by AVR to get unique time stamp. The stamp changed every minute, which was sufficient enough for most applications. Together with unique identifier, the time stamp formed a message to be digitally signed and broadcasted. The signing was realized by means of asymmetric cryptography algorithms and a private key stored in the memory. The encrypted data were then available to read and decrypt using a public key, in turn accessible from an external database, as correlated with the beacon identifier [19].

As it may be seen, the broadcasted and encrypted data make it impossible to perform the "record and play" attack, when someone simply collects the broadcasted info at certain place and time and further present it as authorization data for another place/time. As mentioned before, for untrusted beacons such an attack is trivial with nowadays hardware and software, e.g., using a typical smartphone. As a result, in the case of our proposal a level of trust for broadcasted information is substantially increased. However, the received information still may be exchanged with some other devices, to simulate the presence at given place/time. To restrict such way of cheating, the application receiving the broadcasted messages in turn should be digitally signed and installed in the trusted way. By linking the trusted broadcast and the trusted application receiving this broadcast, one may prove the fact that the receiving device was really present at given place/time, able to receive and memorize the broadcasted information there.

The most important problem related with the implementation of a trusted beacon is related to the efficient implementation of the encryption. As for our tests we assumed an application of AVR processors and cheap modules such as Arduino microcontrollers, in general we had three possible ways to implement the encryption algorithms:

1/ directly in AVR code (C/C++ and assembly language),
2/ by means of specialized hardware modules, directly connected to one of AVR communication links,
3/ as a Java-based library to be executed at PC side and contacted via AVR communication links.

The first way, according to the very limited hardware/software resources of small microcontrollers, seems to be impractical. Even if we successfully implemented and tested several encryption algorithms (not only asymmetric-cryptography algorithms, also e.g., TEA symmetric-cryptography algorithm for assuring secrecy and privacy [20]), we found that the amount of RAM and ROM memory is not sufficient to include, apart the encryption algorithm, also some additional code to deal with some other tasks (such as timings, broadcasting, BLE transmission, etc.). Remember that AVR-based solutions are not based on underlying operating system, thus all the tasks to be performed by the node must be included directly in the program code, starting from such simple procedures as blinking a LED, and finishing on bit-by-bit serial communication and encryption algorithms.

Then, we began to look for some specialized modules to perform the encryption at hardware level. We discovered several proposals, among these the CryptoShield with ATSHA204, HMAC256 and ATECC108 chips seemed to be very promising [21]. Together with specialized chips for traditional RSA encryption and Real-Time Clock RTC add-on, and the popular I²C communication link, the module at the very first view was ideal for our goals. However, we soon detected several drawbacks, making it questionable to apply the module for beacon network. First, the module consumes a lot of energy, even if not used frequently, due to the fact the module is equipped with its own microcontroller (Atmel Trusted Platform Module for CryptoShield version, and ATmega328 for CryptoCape). Second, the size of the module is ideal for basic Arduino controllers such as Uno and Mega, but seems to be much too big in case of the smallest controllers to be used for beacon implementation, not to say about the need of additional connectors to map the pin-out. Third, even if advertised as compatible with the smallest 8-bit controllers, the module in the matter of fact is efficiently working only with the most powerful Arduino boards (32-bits CPUs and Linux-based control), in turn increasing energy consumption and limiting battery life.

V. Architecture of a Testbed

To bypass the implementation problems mentioned in the previous section, we decided to slightly change the architecture of the whole beacon network. We went to the idea of a single cryptography unit to work for several

beacons, with autonomous power supply and trusted communication links with the beacons (Fig. 3). For the initial test, we applied standard wired connections (serial transmission in UART mode, at 115200 Bd). The wired connections were also used to provide power supply for the beacons. We see that the wired solution is certainly not the target one, but as for testing and validating the whole system this solution is much better than a set of wireless connections, fixing (1) the problem of code updates – via serial/USB links, (2) the problem of power supply, and (3) the problem of radio-transmission conflicts and errors.

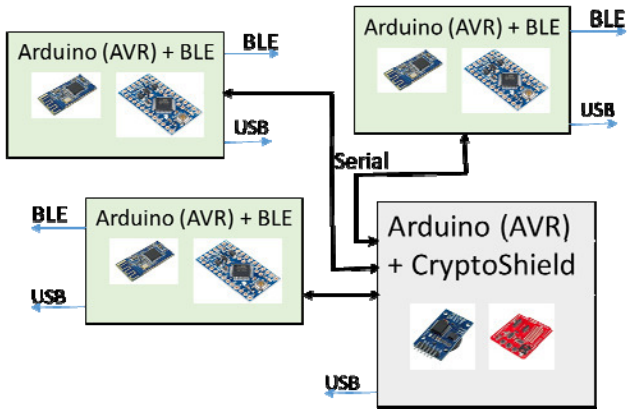


Fig.3. Architecture of the beacon network with single point of encryption implemented as specialized hardware module

The cryptography unit served not only for the pure cryptographic purposes, it also acted as a clock synchronizer (providing RTC data) and activity monitor – as all the beacons periodically quoted for encrypted time stamps, it was quite natural to collect these requests and to report them elsewhere.

Once we wanted to test the timings for the preparation of the encrypted data, and to compare several cryptographic algorithms, we found that the cryptographic module is hardly applied to this goal, mainly due to the limited re-programming possibilities. Thus, we again change the architecture of the testbed, replacing the cryptographic module with a PC serving as a central cryptographic server (Fig. 4). The serial connections to beacons were replaced by standard USB links, and the PC also served as the power supply for the beacons. The server was implemented in Java, with a help of Java Serial Connectivity (JSSC) library and certain cryptographic libraries, including ECC algorithm. Then, we were able to test several implementations of ECC-based encryption, not to say about detailed measurement of timings and the estimation for global energy consumption.

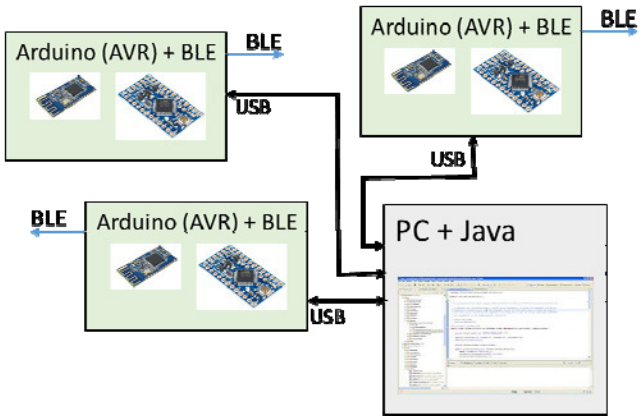


Fig.4. Architecture of the beacon network with single point of encryption implemented as PC-based server

VI. Application Level

Finally, we intended to test the possibility to directly access and monitor the beacons via BLE transmission to be controlled from a portable PC, the solution which is much more practical and efficient than testing on limited Android or Apple devices. Unfortunately, there is no BLE support for a PC running Windows, even if the hardware is there. Thus, to this goal, we used AT mode of the newest BLE-to-USB micro converters, namely HM-10 module [22]. With this module, it is possible to enable so called "directory" mode to detect all BLE modules in the neighborhood. The HM-10 module was directly connected to a USB port of a PC, and the AT commands were applied for BLE monitoring and the detection of signal-strength info. The collected data were processed by a Java program. As a result, it was possible not only to validate the network of beacons, but also to undertake some tests for beacon visibility, possible distortions (due to weather, crowd, other electronic devices in the closed neighborhood, etc.), and many more. Note that, even if BLE transmission is implemented for the newest Android-based devices and Apple phones/tables, the monitoring software is usually very restricted and mainly aims in displaying the most probable location of a beacon at the screen. Our monitoring system provides much more data and may be used for the comparison of different situations, cases, places, etc., not to say about the verification of trust based on encrypted broadcast from the beacons.

HM-10 module was also found to be well adjusted to the communication with Arduino and other AVR-based controllers – this module was finally applied for the tests as a basic BLE unit for each trusted beacon. Here we have to state once again that any of the existing hardware solutions for beacons could not be applied, as there is not a single possibility to change the broadcasted data cyclically (e.g., every second) for a traditional beacon. Again, we applied a combination of AT commands and BLE broadcasting, controlled by underlying Arduino board, to convert the HC-10 module into an efficient beacon simulator.

II. Final conclusions and future work

The testbed architecture described above is the very first step towards an implementation of the target trusted beacon network. As for the future work, we plan:

1/ to measure the efficiency of several encryption algorithms,
2/ to estimate energy consumption due to the encryption tasks,
3/ to test the possibility to provide better cryptographic algorithms and chips to be included in the beacons, especially to find the optimum parameters for ECC encryption algorithm,
4/ to eliminate the need for cable connections, replacing them e.g., by periodic BlueTooth transmission in a safe master/slave mode with traditional pairing,
5/ to validate some other BLE communication modules, as soon as they are available on the market (so far, only HC-10 module seems to be useful for both beacon simulation and detection).

We must also state the fact that the approach described in this paper may be applied to any sensor/actuator network, not only for geolocation beacons, to increase the level of trust for broadcasted messages. Although BLE marketing channel and broadcast are rarely used towards this goal, this is a generic approach that is potentially of great interest for the designers of sensor networks, networks for "intelligent" places and buildings, including "smart cities"[23], applications of Internet of Things [24] and Services [25], and many more.

The results of our work towards trusted geolocalization beacons resulted in a PL/EU/US patent application entitled "Trusted geolocation beacon and a method for operating a trusted geolocation beacon", currently at early-registration phase.

References

[1] What is Bluetooth technology?, http://www.bluetooth.com/Pages/what-is-bluetooth-technology.aspx, 2016
[2] Bluetooth Low Energy Technology, https://www.bluetooth.com/what-is-bluetooth-technology/bluetooth-technology-basics/low-energy, 2015
[3] Beacon Tech Overview, Estimote documentation, http://developer.estimote.com online: 21.10.2015
[4] Bluetooth® Low Energy Beacons, Texas Instruments materials, http://www.ti.com/lit/an/swra475/swra475.pdf, 2015
[5] P. Boddupalli, F. Al-Bin-Ali, N. Davies, A. Friday, O. Storz and M. Wu, Payment support in ubiquitous computing environments, in: Mobile Computing Systems and Applications, proceedings of Fifth IEEE Workshop on. IEEE, 2003
[6] Roberts, P. F. Internet of Things Demands New Social Contract To Protect Privacy. https://securityledger.com/2013/09/internet-of-things-will-force-choice-between-privacy-control/, 2013
[7] Atmel AVR 8-bit and 32-bit Microcontrollers, Atmel documentation, http://www.atmel.com/products/microcontrollers/avr/, 2016
[8] iOS: iBeacon technology overview, Apple documentation, https://support.apple.com/pl-pl/HT202880, 2015
[9] AltBeacon − The Open and Interoperable Proximity Beacon Specification, http://altbeacon.org, 2015
[10] Mark up the world using beacons, Google documentation, https://developers.google.com/beacons/, 2016
[11] Elliptical Curve Cryptography (ECC) Definition, TechTarget reports, http://searchsecurity.techtarget.com/definition/elliptical-curve-cryptography, 2015
[12] RSA cryptosystem, from Wikipedia, https://en.wikipedia.org/wiki/RSA_(cryptosystem), 2016
[13] G. S. Quirino, A. R. L. Ribeiro and E. D. Moreno, Asymmetric Encryption in Wireless Sensor Networks, http://www.intechopen.com/books/wireless-sensor-networks-technology-and-protocols/asymmetric-encryption-in-wireless-sensor-networks - comparison of PKI algorithms and timings, 2016
[14] Adams, C., & Lloyd, S. Understanding PKI: concepts, standards, and deployment considerations. Addison-Wesley Professional, ISBN 978-0-672-32391-1, 11–15, 2003
[15] L. B. Oliveira, D. Aranha, E. Morais, F. Daguano, J. Lopez, and R. Dahab, Identity-Based Encryption for Sensor Networks https://eprint.iacr.org/2007/020.pdf, 2016
[16] Willey W.D., Device Authentication in a PKI, US Patent 8,661,256, 2014
[17] Troxler R.E., Methods, Systems and Computer Program Products for Locating and Tracking Objects, US Patent Application US 2015/0088452, 2015
[18] Balfanz D., Lopes C., Smetters D., Stewart P., Wong H.C., Systems and Methods for Authenticating Communication in a Network Medium, US Patent US 8,156,337, 2012
[19] Rykowski, J., and M. Nomańczuk, Geolocalization beacons – a new way of position determination inside buildings, in: Drives and Control, vol. 12 (200) , Druk-Art. Press, 2015 (in Polish).
[20] Tiny Encryption Algorithm (TEA), from Wikipedia, https://en.wikipedia.org/wiki/Tiny_Encryption_Algorithm, 2016
[21] SparkFun CryptoShield, https://www.sparkfun.com/products/13183, 2016
[22] HM-10 Bluetooth module datasheet, https://www.seeedstudio.com/wiki/images/c/cd/Bluetooth4_en.pdf, 2016
[23] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. A. Pardo, H. J. Scholl (2012), Understanding Smart Cities: An Integrative Framework, proc. of 45th Hawaii International Conference on System Sciences, DOI 10.1109/HICSS.2012.615, 2014
[24] Atzori, L., Iera, A., & Morabito, G. The Internet of Things: A survey. Computer Networks 54 (15), 2787-2805, 2010
[25] Towards the Internet of Services, CORDIS Software & Service Architectures and Infrastructures, http://cordis.europa.eu/fp7/ict/ssai/home_en.html, 2015

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference area 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'16—7th International Workshop on Advances in Business ICT
- AITM'16—14th Conference on Advanced Information Technologies for Management
- ISM'16 - 11th Conference on Information Systems Management
- KAM'16—22nd Conference on Knowledge Acquisition and Management
- UHH'16 - 2nd International Workshop on Ubiquitous Home Healthcare

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 177–180
DOI: 10.15439/2016F461
ACSIS, Vol. 9. ISSN 2300-5963

Foundation for Modular Cloud Logistics

Michael Glöckner[1,2], Björn Schwarzbach[1], André Ludwig[2], and Bogdan Franczyk[1,3]
[1]Leipzig University, Germany
[2]Kühne Logistics University, Hamburg, Germany
[3]Wrocław University of Economics, Poland
{gloeckner | schwarzbach | franczyk}@wifa.uni-leipzig.de - andre.ludwig@the-klu.org

***Abstract*—Logistics service providers are facing the need to collaborate because of ongoing trends like outsourcing and concentration on core competencies. The paradigm of cloud logistics promises to make collaboration in the logistics domain more easy and more flexible by adopting the principles of cloud computing to the logistics domain. To ensure the success of cloud computing to the paradigm of cloud logistics as well, standards have to be developed. With the help of the service blueprinting method, a basic logistics module is developed in order to lay a foundation for cloud logistics.**

I. Introduction

LOGISTICS is affected by increasing outsourcing and thus a concentration on core competencies [1]. Hence, to create sophisticated supply chains and meet complex customer demands, the specialized logistics service providers (LSP) are compelled to work together and collaborate. Flexibility is a driving factor for the selection and orchestration of LSP [2], [3], [4]. Within this context, different paradigms arise and can be adopted to the logistics domain, i.e. service oriented architecture (SOA) paradigm [5], [6] on the one hand and on the other hand the principles of cloud computing (CC) [7], [8]. On the one hand this comprises encapsulation, composability, loose coupling, and reusability (adapted from SOA) and on the other hand virtualization, ad-hoc reconfiguration and interconnectability of resources (adapted from CC). The adoption of cloud principles to the logistics domain to the most possible extent leads to the idea of *cloud logistics* presented in the discussionpaper of [9].

Cloud logistics (CL) is still a topic in its infancy but bears a lot of potential for the digitization of especially small and medium sized enterprises (SME) in the logistics industry. However, those logistics SME 'refuse' to adopt new technologies [10]. As CC is successfully adopted in a genral way in industry, the adoption of its success factors are a promising way to foster CL. Built upon certain standards of descriptions and interfaces, the success of CC relies on compatibility and a high usability for customers as background processes are kept hidden and resources (i.e. infrastructure, platform, software) are offered 'as-a-Service' to the customer. Consequently, providers are compatible, ease of use is enabled, costs can be reduced and thus the main objectives of customers are met in order to change their IT to cloud technology [11]. To make CL become a success as well, such standards and their advantages are obligatory. Cloud principles are to be adopted to the logistics domain in order to lay the foundation of cloud logistics' success.

The work presented in this paper was funded by the German Federal Ministry of Education and Research within the project *Logistik Service Engineering und Management* (LSEM). More information can be found under the reference BMBF 03IPT504X and on the website www.lsem.de.

Delfmann and Jaekel [9] outline a comprehensive service model based on logistics resources and ensuring compatibility through standardized (data) interfaces as a promising field of research for CL. A basic modular approach for CL could standardize interfaces for a sound flow within supply chains.

The paper answers the question: *How can a foundation for cloud logistics be laid in terms of a modular approach?* with the help of the following research questions:

- RQ_1: What are suitable service engineering methods for creating generic service modules?
- RQ_2: What is an appropriate basic concept for modular cloud logistics?

The contribution of the paper is a basic modular logistics concept in order to enable cloud logistics. In section II background knowledge is given and the reader is briefly introduced to the basics of cloud logistics and service blueprinting, which is the used method for service engineering. Section III presents the basic module concept of cloud logistics. The paper is concluded by a summary and an outlook in section IV.

II. Background

A. Method

As cloud logistics is currently a theoretical concept [9], empirical observations are not possible. Thus, a procedure based on the design-science paradigm for information systems [12] is applied. The analysis is conducted during the current section, where the basic principles of cloud logistics are described and later taken as the starting point for the design phase. As cloud logistics focuses especially on the flow of physical as well as informational entities, concepts and methods of product service systems, *service blueprinting* in particular, are invoked during the design phase. [13] deliver a seminal work reviewing the "State of the art and research challenges of Product-Service Systems Engineering". Further, the method of domain engineering [14] is taken into account in order to find 'common points' and 'varying points', and to create *configurable* requirements and architecture of the logistics domain.

B. Cloud Logistics

Delfmann and Jaekel present the basic idea of cloud logistics (CL) in their discussion paper about the logistics for the future [9]. They cite the German Logistics Association (BVL 2012) [15] and define CL as:

An environment of 'virtual' systems that facilitate supply chains' overall coordination and use of distributed resources, capacities, processes, and services from supply chain partners. These systems are based on advanced information and communication technologies that leverage modern Internet services.

Delfmann and Jaekel mention that the adoption of the cloud principles to other domains does not mean to implement cloud computing solutions in the target domain (however, cloud computing indeed helps to manage and support cloud principles in terms of information systems technology). Rather they emphasize explicitly that the adoption of cloud principles to another domain aims at interpreting the domain itself with the help of a cloud paradigm. With regards to the definition of cloud computing [7], [8], this comprises a pool of virtualized resources, which are easily usable and accessible. The resources offer the possibility of a dynamic (re-)configuration and an adjustment to differing workload. By using same the resources for several customers, an optimal resource allocation and utilization can be achieved. The resources of the network are accessible on demand with minimal management effort 'as-a-Service'.

The interpretation of cloud principles in the logistics domain leads to the paradigm of 'cloud logistics', 'Logistics-as-a-Service' (LaaS) [16], or 'Supply Chain as a Service' (SCaaS) [17]. Core of the idea is to define a new service type next to Infrastructure-as-a-Service, Platform-as-a-Service and Software-as-a-Service. Involved resources of logistics are hidden behind a common predefined interface and can be added, reconfigured and removed 'on-demand' responding to customers' needs [17]. However, reaction times are of course higher, as resource allocation (e.g. transferring trucks to pick up goods) implies the overcoming of physical distances. As competitors now offer more and more homogeneous services (e.g. transportation, warehousing, packing, track and trace), outsourcing is more and more encouraged [18], as seen in other industries before [19]. This process makes it possible to turn fix costs into variable costs [20], as needed logistics services (e.g. transportation, picking) and resources (e.g. trucks, warehouses) can be ordered on demand and thus, logistics assets do not have to be owned directly. Service levels are also well known in logistics [21].

So, just old wine in new skins? Of course outsourcing and insourcing of capabilities, processes and resources is nothing new to logistics. However, with a shift from being rather closed and striving for competitiveness only by low prices, LSP should understand the disruptive paradigms of the cloud and digitization with their inherent transparency and flexibility as a chance. With the adoption of this mindset, the focus could be transferred to an increasing specialization and providing higher service quality in order to be successful on the market.

C. Service Blueprinting

Whereas services in logistics are traditionally related to classical business service on the one hand, services in CC on the other hand are related to electronic services. In order to establish a cloud logistics paradigm, real-world business service and electronic service have to be combined. Usage of service engineering methods for product-service systems (a comprehensive overview can be found in [13]) seems to be appropriate. The method of service blueprinting [23], especially the modified version *extended service blueprinting* by [22], offers suitable aspects to describe services that are based on both business services electronic services. Figure 1 presents the extended service blueprinting using BPMN.

Hara et al. [22] distinguish between a *behavior blueprint* that represents the machines and their related software involved in a service (= electronic services) as well as an *activity blueprint* that represents the 'humanware' and their related supporting software (= business services). General depiction method is the business process management notation (BPMN) in order to ensure a common and easily understandable communication standard. Services in general are seen as a set of functions that have a possible value for customers in terms of changing one or more receiver state parameters (RSP) [24], [25]. Those RSP could be structured down to the lowest level where they represent basic functions and are mapped afterward to specific process steps of services in order to highlight their importance in the context of interaction with the customer. The change of the RSP is goal of business activities and thus they form the input, output and requirements from the stakeholders' sides. Further, two important lines are introduced: the line of interaction (separating service consumers and service providers) and the line of visibility (separating 'onstage (visible)' and 'backstage (invisible)' activities performed by the provider). An inter-relation between the activity blueprint (humanware + related software) and the behavior blueprint (machines + related software) is obligatory for the extended service blueprinting. A connection is to be established between the functions of the RSP and the appropriate process steps.

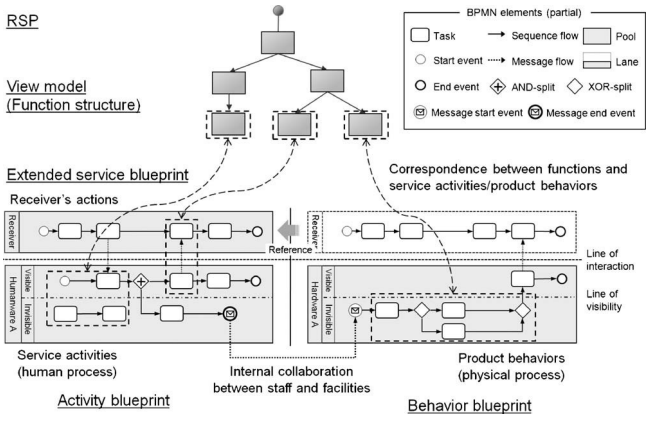


Fig. 1. Notation of the extended service blueprint taken from Hara et al. [22].

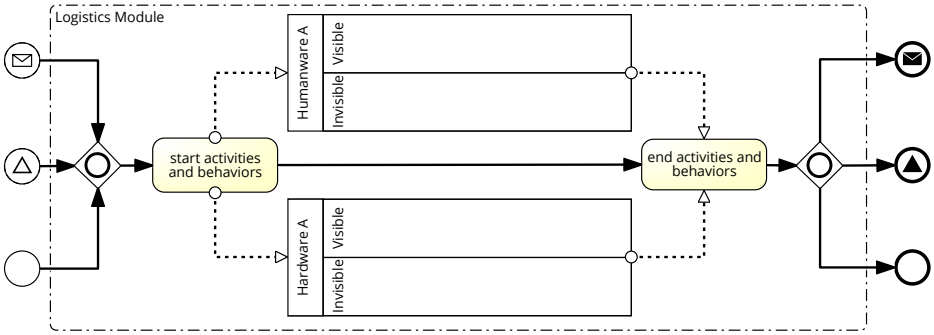


Fig. 2. The extended service blueprint is encapsulated with logistics interfaces in order to create the basic logistics module and enable cloud logistics.

III. Basic Logistics Module

A. Transferring the Extended Service Blueprinting to Logistics

When creating a basic logistics module, which is based on the extended service blueprinting, the concept shown in Figure 1 has to be encapsulated and logistics characteristics have to be taken into account. From the basic logistics module, specific logistics services can be derived that incorporate distinct logistics resources in order to fulfill logistics functions. Because of the modules with common interfaces, those logistics functions can be combined and thus cloud logistics is enabled.

Common points of logistics are the general planning, operating and controlling of both the flow of goods and the flow of information [26]. The encapsulation is depicted in Figure 2. The activity blueprint (aka. Service activities or human process) as well as the behavior blueprint (aka. product behavior or physical process) are triggered by events. Those events are in every case a distinct trigger-signal (interface middle-event) from the orchestration of the central integrator and the required information (interface top-event) from the central integrator as well as from the previous logistics module. Further, the goods that are involved in the logistics service and their physical allocation (interface bottom-event) are an important input for every logistics module. Those two kinds of events form the common points of the logistics domain. However, cloud logistics is information intensive and comprises also information-centric services (e.g. customs clearance, identification or track and trace). Hence, as the allocation of physical goods is not always obligatory for information-centric services, an inclusive gateway is chosen to bundle the input flows, or to distribute the output flows, respectively. This is a varying point for services of the logistics domain. With the final output of information, parameters of service quality as well as the trigger-signal for the next logistics module (if existant) can be forwarded. The flow of value in terms of financial aspects is also involved in logistics but not explicitly depicted. On the one hand it is not in the main focus of logistics, and on the other hand it could be regarded as a kind of informational flow in the context of online banking (even though, there are much higher formal and security requirements).

The RSP are the connection to several stakeholders (logistics integrator, logistics service providers and customers) and have to be taken into account as they are goal, input, output and requirements. With the standard of the interface containing three events as input and output, several logistics modules can be put into a (complex) chain in order to connect several logistics resources enabling cloud logistics.

B. Example

Typical activities of contract logistics are taken from [2]. The rough example comprises a transportation process involving different modes of transport that are carrying hazardous goods, as depicted in Figure 3. Hence, four different logistics modules are to be combined, whereas each module could be sourced from a different LSP. The highest priority is the qualification of having the permission of *Hazardous Goods Management*. This implies next to specific licenses the ability of monitoring and controlling those hazardous goods. The different modes of transport require an overarching module of *Intermodality Management*. Its function is the initiation, monitoring and controlling of different modes of transport as well as managing the handling processes between them. Furthermore, the two modes of transport are each invoked through a logistics module. The *Rail Transportation* is the long-distance run. After finishing, the hazardous goods are

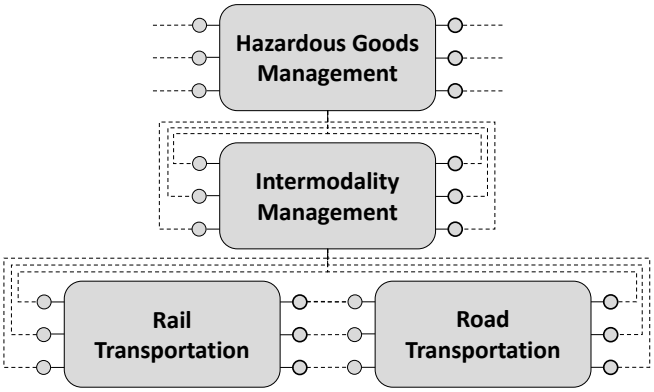


Fig. 3. Logistics Modules of basic logistics functions are combined and invoke other modules in order to develop and operate a complex logistics service.

loaded from the train to a truck and the *Road Transportation* is conducted for the short-distance run on the last-mile.

The highest priority module (hazardous goods management) is the initial point for the complex logistics service. During planning phase the connection to the other modules is established and RSP are presented to the further invoked modules as input parameters and requirements. This invocation is either done automatically by the hardware lane of the module or manually by the humanware lane (see Figure 2).

IV. Conclusion

In the paper, a brief introduction to cloud logistics is given and the need for standardized logistics modules as an enabler is outlined. In order to answer research question 1 the method of extended service blueprinting is presented as a suitable possibility for designing standardized logistics modules as it enables both human and machine interaction and meeting logistics characteristics of informational and physical flows. In order to answer research question 2, a basic concept for a logistics service module is presented and applied to a rough example of a complex logistics service. Summarizing, a foundation for cloud logistics is laid in terms of a basic logistics modular approach concept. The modular approach with standardized interfaces enables the access to resources of the logistics network on demand with minimal management effort 'as-a-Service'.

Implication for the scientific community on the one hand is an initial concept of module in order to enable cloud logistics. On the other hand implication for business is to introduce the cloud logistics paradigm as beinng a disruptive force, bringing a high potential for increased flexibility and efficiency for collaborative complex logistics services.

Future fields of research should address the module itself in terms of further developing its details such as the interfaces, formats of information interchange and invoking techniques for nesting and embedding. Another field of interest could be the standardization of the internal logistics transformations of the modules (e.g. [26] describe logistics as a transformation of goods in space and time). With this module-internal standardization, the standardization of the interfaces as well as a categorization of modules can be realized. Prioritization of different modules, as shown in the example, should follow distinct rules, which form another field of future research.

References

[1] J. Langley and M. Long, "2016 third-party logistics study: The state of logistics outsourcing: Results and findings of the 20th annual study," vol. 20, 2016.

[2] A. Aguezzoul, "Third-party logistics selection problem: A literature review on criteria and methods," *Omega*, vol. 49, pp. 69–78, 2014. DOI: 10.1016/j.omega.2014.05.009.

[3] T. Solakivi, J. Töyli, and L. Ojala, "Logistics outsourcing, its motives and the level of logistics costs in manufacturing and trading companies operating in finland," *Production Planning & Control*, vol. 24, no. 4-5, pp. 388–398, 2013. DOI: 10.1080/09537287.2011.648490.

[4] R. Wilding and R. Juriado, "Customer perceptions on logistics outsourcing in the european consumer goods industry," *International Journal of Physical Distribution & Logistics Management*, vol. 34, no. 8, pp. 628–644, 2004. DOI: 10.1108/09600030410557767.

[5] A. Arsanjani, G. Booch, T. Boubez, P. Brown, D. Chappell, J. deVadoss, T. Erl, N. Josuttis, D. Krafzig, M. Little, B. Loesgen, A. T. Manes, J. McKendrick, S. Ross-Talbot, S. Tilkov, C. Utschig-Utschig, and H. Wilhelmsen. (2009). Soa manifesto, [Online]. Available: http://www.soa-manifesto.org/SOA_Manifesto.pdf.

[6] T. Erl, *SOA: Principles of service design*, 5. print, ser. The Prentice Hall service-oriented computing series from Thomas Erl. Upper Saddle River, NJ [u.a.]: Prentice Hall, 2009, ISBN: 9780132344821.

[7] P. Mell and T. Grance, "The nist definition of cloud computing," *Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology Gaithersburg*, 2011.

[8] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, p. 50, 2008. DOI: 10.1145/1496091.1496100.

[9] W. Delfmann and F. Jaekel, *The cloud - logistics for the future? discussionpaper*, German Logistics Association - BVL International, Ed., 2012.

[10] U. Arnold, J. Oberlander, and B. Schwarzbach, "Logical—development of cloud computing platforms and tools for logistics hubs and communities," in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*, 2012, pp. 1083–1090.

[11] P. Gupta, A. Seetharaman, and J. R. Raj, "The usage and adoption of cloud computing by small and medium businesses," *International Journal of Information Management*, vol. 33, no. 5, pp. 861–874, 2013. DOI: 10.1016/j.ijinfomgt.2013.07.001.

[12] A. Hevner, S. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[13] S. Cavalieri and G. Pezzotta, "Product–service systems engineering - state of the art and research challenges," *Computers in Industry*, vol. 63, no. 4, pp. 278–288, 2012. DOI: 10.1016/j.compind.2012.02.006.

[14] K. Czarnecki and U. Eisenecker, *Generative programming: Methods, tools, and applications*. Boston: Addison Wesley, 2000, ISBN: 0-201-30977-7.

[15] German Logistics Association. (2012). 6th international scientific symposium on logistics: Call for papers, [Online]. Available: http://www.bvl.de/misc/filePush.php?id=15746&name=ISSL12+Call+for+Papers.

[16] K. Klingebiel and A. Wagenitz, "An introduction to logistics as a service," in *Efficiency and logistics*, ser. Lecture Notes in Logistics, U. Clausen, M. ten Hompel, and M. Klumpp, Eds., Springer, 2013, pp. 209–216, ISBN: 978-3-642-32837-4. DOI: 10.1007/978-3-642-32838-1_22.

[17] J. Leukel, S. Kirn, and T. Schlegel, "Supply chain as a service: A cloud perspective on supply chain systems," *IEEE Systems Journal*, vol. 5, no. 1, pp. 16–27, 2011. DOI: 10.1109/JSYST.2010.2100197.

[18] T. H. Davenport, "The coming commoditization of processes," *Harvard business review*, vol. 83, no. 6, pp. 100–108, 2005.

[19] M. Reimann, O. Schilke, and J. S. Thomas, "Toward an understanding of industry commoditization: Its nature and role in evolving marketing competition," *International Journal of Research in Marketing*, vol. 27, no. 2, pp. 188–197, 2010. DOI: 10.1016/j.ijresmar.2009.10.001.

[20] P. Matthyssens and K. Vandenbempt, "Moving from basic offerings to value-added solutions: Strategies, barriers and alignment," *Industrial Marketing Management*, vol. 37, no. 3, pp. 316–328, 2008. DOI: 10.1016/j.indmarman.2007.07.008.

[21] J. T. Mentzer, D. J. Flint, and J. L. Kent, "Developing a logistics service quality scale," *Journal of Business Logistics*, vol. 20, no. 1, p. 9, 1999.

[22] T. Hara, T. Arai, Y. Shimomura, and T. Sakao, "Service cad system to integrate product and human activity for total value," *CIRP Journal of Manufacturing Science and Technology*, vol. 1, no. 4, pp. 262–271, 2009. DOI: 10.1016/j.cirpj.2009.06.002.

[23] L. G. Shostack, "How to design a service," *European Journal of Marketing*, vol. 16, no. 1, pp. 49–63, 1982.

[24] T. Sakao and Y. Shimomura, "Service engineering: A novel engineering discipline for producers to increase value combining service and product," *Journal of Cleaner Production*, vol. 15, no. 6, pp. 590–604, 2007. DOI: 10.1016/j.jclepro.2006.05.015.

[25] T. Arai and Y. Shimomura, "Proposal of service cad system - a tool for service engineering -," *CIRP Annals - Manufacturing Technology*, vol. 53, no. 1, pp. 397–400, 2004. DOI: 10.1016/S0007-8506(07)60725-2.

[26] T. Gudehus and H. Kotzab, *Comprehensive Logistics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ISBN: 978-3-642-24366-0. DOI: 10.1007/978-3-642-24367-7.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 181–188
DOI: 10.15439/2016F385
ACSIS, Vol. 9. ISSN 2300-5963

An Analysis of CEN/WS BII

Veit Jahns
University of Duisburg-Essen, Institute for Computer Science and Business Information Systems,
Universitätsstraße 9, 45117 Essen
Email: veit.jahns@icb.uni-due.de

Frank-Dieter Dorloff
University of Duisburg-Essen, Institute for Computer Science and Business Information Systems,
Universitätsstraße 9, 45117 Essen
Email: frank.dorloff@icb.uni-due.de

***Abstract*—This position paper discusses the outcome of the standardization initiative CEN/WS BII aiming to support the implementation of a complete end-to-end e-procurement process, that may be used in cross-border business all over Europe and in particular in the public sector. The discussion includes the background of this initiative, its approach, and its outcomes and shall help to understand how the CEN/WS BII worked and if and how these deliverables are appropriate to fulfill the aim of harmonized European-wide e-procurement in the public sector. Furthermore, based on the discussion, implications for future research in Information Systems with respect to standardization of e-procurement and interoperability of information systems are provided.**

I. Introduction

SINCE the rise of e-government, modernizing the public administration in Europe and enhancing its efficiency is one of the major drivers for introducing information technology on every level of public administration. Especially in the field of public e-procurement, the European Commission expects positive effects like seamless and quicker processes and savings of around €100 billion [1].

A crucial prerequisite to reach these goals is a fully end-to-end e-procurement, including the processes of notification, tendering, ordering, fulfillment, and invoicing [2]. This includes that all relevant stakeholders (buyers, sellers, and service providers) participating in public Business have to exchange their business documents automatically based on Europe-wide used and accepted standards for processes, documents, and rules, i.e., the interacting in-formation systems have to be interoperable.

Nowadays, exchanging e-procurement documents in Europe between the various stakeholders is still an unsolved challenge that has been addressed in various research works. For exchanging such documents electronically, all relevant stakeholders have to agree on accepted exchange rules and standards. In the early phase of e-procurement, the main challenge was a lack of appropriate e-procurement- standards to be used. Nowadays and contrary, the problem is the multitude and heterogeneity of available standards stakeholders can choose from or have chosen already. In consequence, in the various member states of the European Union individual e-procurement systems got established which may be interoperable within the corresponding member state, but are not interoperable cross-border, i.e. with the e-procurement systems of the other member states.

For instance, in invoicing the Scandinavian member states base their approach on the UBL Invoice for exchanging invoices, e.g., Svefaktura [3] and EHF Invoice [4], while other member states base their national invoice standard on the UN/CEFACT Cross-Industry Invoice (CII) like it is done in Germany with the ZUGFeRD invoice [5].

One very restrictive solution of this challenge would be that all stakeholders have to agree on only one possibly new developed e-invoice standard to be used mandatory in the public sector of each European member state. But such a solution may have crucial drawbacks on the economic, technical and political level with winners and losers. If, for instance, only the UBL Invoice would be accepted as the one and mandatory standard then all those member states using e-procurement systems based on UN/CEFACT CII will lose their e-procurement investments because they have to implement UBL-based e-procurement systems. In consequence, everyone would try to protect its e-procurement investments by promoting their own standard and will abandon other standards—a scenario what in the context of information systems interoperability literature was described as “empire building” [6]. By this, standards will become an instrument to impose the own preferences on the other stakeholders.

How can such an empire building mentality be avoided? The nature of standardization processes has been addressed in many research works, e.g., [7]-[10]. Besides these research works, this issue has been also addressed in non-scientific projects like the project CEN workshop on Business Interoperability Interfaces (CEN/WS BII) by the European Committee for Standardization (CEN). As a pre-standardization initiative it was concerned to specify and harmonize the manifold requirements on a public e-procurement as well as to give guidance how existing e-procurement standards—including competing e-procurement standards—can be used to implement a European-wide public e-procurement.

In this position paper it is argued that the approach chosen by the CEN/WS BII is a way to deal with the aforementioned empire building mentality. The focus was on requirements and on giving guidance how to use existing e-procurement standards, no existing implementation was preferred. Instead,

The participation of the authors in CEN/WS BII was supported by the German Association Supply Chain Management, Purchasing and Logistics (BME).

it allowed each stakeholder in Europe to preserve its existing e-procurement systems but also to evolve it so that it becomes interoperable with the e-procurement systems of other stakeholders. The approach to achieve this objective as well as the initiative CEN/WS BII itself will be presented and discussed to outline implications for future research with respect of the standardization and interoperability of e-procurement systems.

The paper is organized as follows: in the following second section, the design of this research is described and in the third section, the background of CEN/WS BII. In the fourth section the approach developed by CEN/WS BII is reconstructed while the fifth section describes the various workshop outcomes and the sixth section discusses the corresponding insights that were gained from the various phases of CEN/WS BII. The limitations of this case study and possible future work will be discussed in the seventh section and, as usual, a conclusion will finish this paper.

II. The Development of CEN WS/BII

A. The first phase

The first phase of CEN/WS BII started 2007 in Copenhagen as a so-called CEN workshop and lasted until 2010. CEN workshops are no formal standardization initiatives, but rather informal groups of individuals and/or organizations giving recommendations for possible standards. Their outcome is defined as a CEN workshop agreement (CWA) reflecting the consensus of the group on a particular issue [11]. Such an issue might be giving guidance on the structure, content and implementation of a standard or specifying the requirements for proposing a new standard.

The purpose of the workshop CEN/WS BII was to find an agreement on how e-business standards used in European states can be merged into a public e-procurement standard accepted throughout Europe. The main focus of the first phase was to give guidance how to use UBL to implement a European public e-procurement. The deliverables of the first phase were published as CWA 16073:2010. This CWA cannot be retrieved from CEN anymore, because CWAs—without prolongation by the workshop—are valid for three years only. However, the CWA is still available at the workshop's website [12].

B. The second phase

The second phase of CEN/WS BII started in 2010 and lasted until the end of 2012. It had a wider focus than the first phase and in addition, it should give guidance how to use UN/CEFACT XML to implement public e-procurement. Furthermore, it provides an advanced methodological foundation by specifying core business requirements and by modeling the semantics of the public e-procurement business transactions which then could be mapped to the semantics embedded in the messages of UBL and UN/CEFACT XML.

The outcome of this phase became more complex and in consequence five CWAs were published covering the architecture of CEN/WS BII (CWA 16558:2013), notification (CWA 16559:2013), tendering processes (CWA 16560:2013), electronic catalogues (CWA 16561:2013), and post-award processes (CWA 16562:2013). These CWAs are still active and can be retrieved from the CEN website.

C. The third phase

The third and final phase of CEN/WS BII started in 2013 and finished its work end of 2015. The outcome was published in five CWAs as well, which have the same structure as the CWA of second phase. There were no major changes in the underlying architecture of CEN/WS BII, but rather refinements and improvements, for instance, by establishing a business term vocabulary. In line with the structure of the second phase, the CWAs were published recently as CWA 17025:2016, CWA 17026:2016, CWA 17027:2016, CWA 17028:2016, and CWA 17029:2016.

III. Participants and stakeholders

The main goal of CEN/WS BII is to provide a guidance to implement an acceptable, efficient, and standardized public e-procurement process throughout Europe and to ensure that preferably all necessary and most relevant business requirements are gathered. Furthermore, CEN/WS BII coordinated its activities with other relevant European and international activities, such as GS1 in Europe, the German standards Xvergabe and BMEcat, the Multi Stakeholder Group of Experts on Public Procurement (EXEP). CEN/TC 434 on electronic invoicing and with international organization like OASIS UBL and UN/CEFACT XML mainly with respect to syntax solutions.

Experts from public authorities, standardization bodies, universities, as well as, software vendors from more than 20 European states and institutions of the European Union joined the workshop meetings regularly. This broad participation aimed to include the widest expertise possible for structuring and developing this upcoming new standard for e-procurement.

The authors of this paper were active participants of CEN/WS BII on behalf of the German Association Supply Chain Management, Purchasing and Logistics (BME) and the University of Duisburg-Essen main developer and maintainer of the German wide used e-catalogue standard BMEcat. As such, the authors were heavily involved in the architecture and in the development of catalogue-related deliverables.

IV. The CEN WS/BII approach

The approach taken by CEN/WS BII is based on three main propositions:

(1) Since XML is the base language for many standards, all these standards share a common ground. This eases the conversion of data structures and documents between different standards, in particular, if the underlying concepts expressed in these standards are the same or at least very similar.

(2) There is more stability on the semantic level than on the syntactic level. Names and sequences of syntax elements may change over time, but the key semantic concepts expressed by the syntax elements usually stay a longer time, for

instance, key concepts like invoice number, VAT rate, address, or net price.

(3) There is a significant amount of requirements that have to be fulfilled in all member states. Identifying and harmonizing these “core” requirements enable the member states to comply and provide a sufficient base that makes the European e-procurement systems in the e-procurement chain (cf. Fig. 1) interoperable.

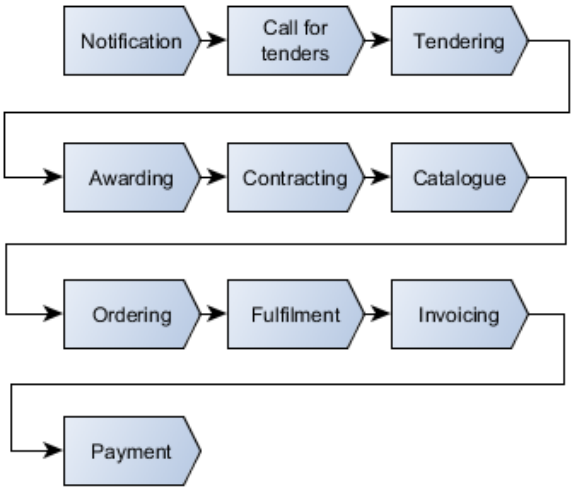


Fig. 1 E-Procurement chain

In line with the proposition (1) and (2), the approach by CEN/WS BII does not include a syntax specification, but rather a guide how to use existing syntax standards like UBL and UN/CEFACT XML and to implement the identified requirements. As there are a lot of different requirements throughout the e-procurement chain, these are bundled by the architecture of CEN/WS BII with the corresponding business processes and information models to so-called a profiles. Fig. 2 shows a simplified extract of the model of the CEN/WS BII architecture describing the concept of a profile and its related concepts.

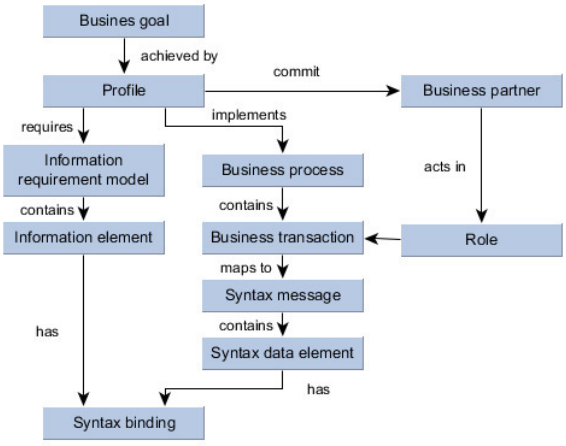


Fig. 2 Simplified model of the architecture of CEN/WS BII

The starting point of a profile is a specific business process aiming at fulfilling a specific business goal. The business process describes a sequence of activities to be performed by partners involved in this business process. The involvement of an actor is described by a so-called role. While acting in their role in the business process, the business partners exchange messages within a transaction, as it is illustrated in Fig. 3.

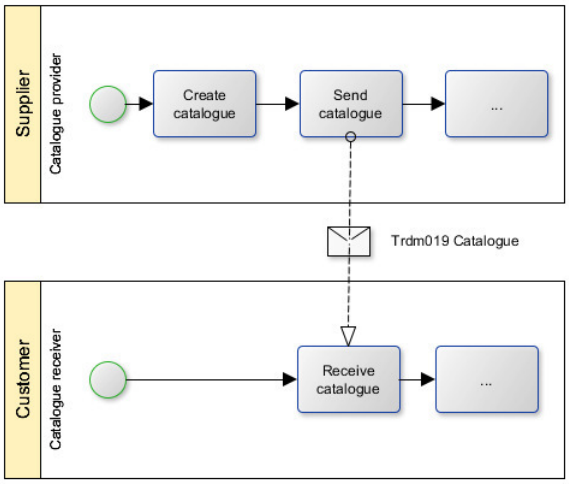


Fig. 3 Extract of the catalogue business process

As an example, an extract of the business process “Catalogue” is shown describing the business process of exchanging catalogues between a supplier and a customer. The business process is modelled using BPMN [13].

Part of each business process is a flow of messages from the supplier—acting in the role of catalogue provider—to the customer—acting in the role of catalogue receiver. In this case, the message sent is the catalogue message containing information on the offered goods and services.

The information in such messages is additionally described in the information requirement model (IRM). This model covers all the information requirements needed to be exchanged within a transaction as part of a business process. The IRMs are specified using the notation of the UML class diagram. Each of the IRMs has a strict hierarchical structure, as it is illustrated in Fig. 4.

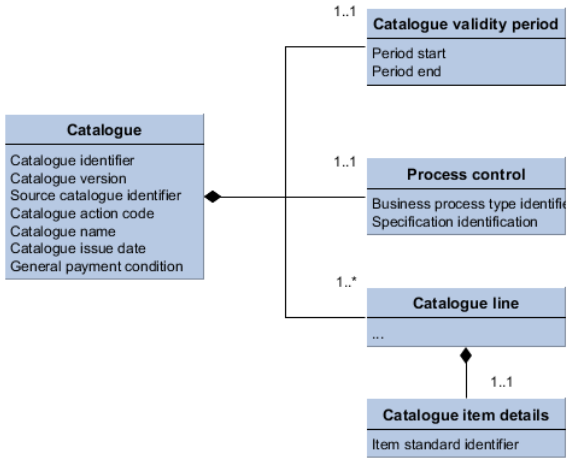


Fig. 4 Extract of the IRM for the transaction “Trdm019 Catalogue”

In IRM shown in Fig. 4 specifies the information exchange in the transaction “Trdm019 Catalogue”. The several classes combine semantically related attributes and represent

semantic parts of the transaction that is why they are related in a composition relationship to each other.

To complete the specification of the information exchange in a transaction, each of the attributes—representing an information requirement—is specified in more detail. Giving an example, in a catalogue process a catalogue has to be exchanged by the business partners. To identify the goods and services listed in the catalogue uniquely, an identifier has to be provided by the so-called information requirement "Item standard identifier".

Table 1 shows an extract of the specification specifying the concept of a unique identifier for good or a service listed as an offer in a catalogue. The definition of an information requirement includes also a specification of the cardinality, the data type, and a mapping to a corresponding business requirement. These were left out for clarity.

TABLE 1.
EXTRACT OF AN INFORMATION REQUIREMENT DEFINITION

InfReqId	Business term	Usage
tir19-092	Item standard identifier	An item identifier based on a registered scheme.

This specification covers only the semantic level. In consequence, the IRM for a transaction only specifies the concepts used in the transaction. In the case of the information requirement "Item standard identifier", this means it is a part of the information requirement “Catalogue” describing the content of a catalogue of goods and services.

To actually exchange a catalogue, a message format is needed that allows "wiring" the catalogue from the supplier to the customer. As CEN/WS BII does not provide such message formats, something else is needed. In this case, existing standards for message formats are used, which are mapped to the information requirements via a so-called syntax binding. In the case of the “Item standard identifier”, the syntax binding to the messages UBL Catalogue and UN/CEFACT XML Cross-industry catalogue by XPath expressions as specified in Table 2.

TABLE 2.
EXAMPLES FOR SYNTAX BINDINGS TO UBL AND UN/CEFACT XML

InfReqId	Syntax binding
tir19-092	Catalogue/cac:CatalogueLine/cac:Item/ cac:StandardItemIdentification/cbc:ID
tir19-092	CrossIndustryCatalogue/ CICHSupplyChainTradeTransaction/ IncludedCICLSupplyChainTradeLineItem/ SpecifiedCatalogueTradeProduct/ID

In this table, using the concepts described in Fig. 2 the information requirement “Item standard identifier” is bound to a specific syntax data element cbc:ID of the syntax message UBL Catalogue as well as to a specific syntax data element ID of the syntax message UN/CEFACT Cross-industry catalogue.

Following the CEN/WS BII approach described before, all business process related aspects of public procurement are specified by the profiles. These profiles cannot cover all possible aspects of public procurement in European states, but they are focused to the “core” aspects, i.e. those aspects that are equal or very similar in the various European states. This is in line with the third proposition mentioned before focusing on the core requirements of public procurement.

V. THE OUTCOME OF CEN/WS BII

A. Overview

The various deliverables of the third phase of CEN/WS BII were published as CEN workshop agreements (CWA) covering all phases of the e-procurement chain (cf. Fig. 1) [14]. The profiles are organized in one general CWA specifying the methodology and architecture of CEN/WS BII and four CWAs covering the e-procurement chain. Each CWA specifies profiles, transactions, and syntax bindings or provides guidelines for specific topics related to the implementation of the profiles and transactions. Table 3 gives an overview on these CWAs.

TABLE 3.
CWA OF THE THIRD PHASE OF CEN/WS BII

CWA	Title	Parts
17025	Methodology and architecture	19
17026	Notification profiles and transactions	11
17027	Tendering profiles and transactions	36
17028	Catalogue profiles and transactions	29
17029	Post-award profiles and transactions	38

B. Methodology and architecture

This CWA covers the methodological and architectural aspects for the other CWA by CEN/WS BII. It describes, how the other CWA are structured, how the business requirements are gathered and described, how the processes and data are modelled, and how the bindings to the various syntaxes are specified, etc.

The two parts 109 and 116 of the CWA are dedicated on the methodology and the architecture. Part 109 elaborates on the concept of core and especially on those core business requirements in public e-procurement that are most relevant for any member state. Part 109 outlines the definition of a core business requirement as well as the approach used by CEN/WS BII to identify these core business requirements.

Part 116 provides a business term vocabulary, which was the base for all the profiles and transactions provided by the four other CWAs. By this business term vocabulary all profiles and transactions are aligned with each other sharing the terms used in the profiles and transactions of all CWAs. This business term vocabulary can be seen as a preliminary ontology of public e-procurement.

C. Notification profiles and transactions

The CWA on notification covers the first phase in the e-procurement chain (cf. Fig. 1). The profiles specified in this CWA are rather specific for public e-procurement and do not cover the special needs of the private sector in the field of e-sourcing. Public administrations have to account for the money spent and they are not allowed to prefer certain sup-

pliers as well as they have to make the process of finding and selecting suppliers fully transparent.

This is reflected in the CWAs by providing profiles and transactions for notifying the public on publishing information on current sourcing activities and their outcome as well as profiles and transaction for searching published notification. As the underlying business processes are very specific for public e-procurement, various directives of the European Union, in particular 2014/23/EU [14], 2014/24/EU [16], and 2014/25/EU [17] published in the course of the third phase of CEN/WS BII, are sources for business requirements addressed by the profiles and transactions of this CWA.

D. Tendering profiles and transactions

The parts of the e-procurement chain covered by the second CWA is also specific for public e-procurement. But while the former CWA addresses the issues related to notifying the public on sourcing activities by contracting authorities, this CWA addresses all issues related to sourcing activities themselves. The parts of this CWA are profiles and transaction for calling for tenders, for receiving tenders, as well as conducting the awarding and contracting the most appropriate tender.

In addition, profiles and transactions are provided by this CWA to provide the accompanying documents needed in the context of public procurement, namely the qualification and the virtual company dossier. The related profiles and transaction cover the processes for evaluating the capabilities of suppliers submitting tenders and for self-declaration by a supplier that all necessary regulatory criteria are met.

As Table 3 indicates, this CWA is the one with the most parts, i.e., with the most profiles and transaction. The number of parts is not driven by the amount of processes covered in this CWA, but rather by the complexity of the covered processes. The complexity has its origin in the multitude of goods and services purchased by public administrations as well as the multitude of suppliers and of public administrations themselves. Public administrations buy almost everything from simple goods for maintenance, repair and operations to complex buildings and machines. A public administration can be a small municipality with little IT capabilities or a national ministry having an advanced IT infrastructure at hand.

To address these multitudes, several profiles and transactions are provided for various maturity levels. For instance, the profile for calling for tenders comes in three shapes. Firstly, a simple call for tenders is provided allowing only the provision of the call for tenders and unstructured documents specifying the goods and services to be tendered, qualification criteria, etc. Secondly, an advanced call for tenders is provided allowing the provision of a structured specification of the goods and services, qualification criteria, etc. Thirdly, the advanced call for tenders can be combined with a so-called pre-award catalogue request allowing the requirements on the requested goods and services in a structured and vendor-neutral way based on classification systems for goods and services.

As the processes covered by this CWA are very specific to the public procurement, regulations are the main source for business requirements, in particular the aforementioned directives by the European Union.

E. Catalogue profiles and transactions

The CWAs discussed two sections before, are located in the pre-award phase of the e-procurement chain. The CWA for catalogue profiles and catalog transactions can be seen as the bridge between the pre-award phase and the post-award phase. Consequently, this CWA provides specifications of core processes and transactions for both e-procurement phases, in particular all core processes and transactions for exchanging catalogues in the tendering phase as well those needed after the awarding of a supplier. Some transactions related to pre-award catalogues and specified in this CWA are even used in the profiles in the CWA on tendering.

In addition, this CWA provides two guidelines as well. One guideline elaborates on the implementation of pre-award catalogues and illustrates by providing examples how the various profiles and transaction can be used.

The other guideline elaborates on the usage of classification systems with the various profiles and transaction. This guideline gives a survey on the four major classification systems CPV, UNSPSC, GS1 GPC and eCl@ss as well as many domain-specific classification systems like ATC, TARIC, ETIM, NCS, or ClaDiMed. Each of the classification systems is described and illustrated as well as examples are given what to do if the classification system is to be used in a transaction. Furthermore, issues of managing and providing classification systems are discussed.

F. Post-award profiles and transactions

The last CWA covers all core profiles and transactions for the post-award phase. These processes and exchanged transactions specify how to place orders (ordering), fulfill orders (fulfilment) as well as send invoices (invoicing) and pay invoices (payment). In the post-award phase, e.g., in ordering, fulfillment, invoicing and payment, the public and private sector are more similar in their goals, business requirements, and activities.

As a consequence, this area is more advanced than the pre-award area, because the standards developed for the private sector can be used in the public sectors as well. In particular, for almost every transaction there are syntax bindings to UBL and UN/CEFACT XML available. Compared to the CWAs from the pre-award area, it is easier to implement syntax messages available. In the case of the pre-award, appropriate syntax messages for a number of transactions have still to be developed by the standardization bodies.

A key profile and transaction in the post-award CWA is the profile for invoicing. In parallel to CEN/WS BII, a technical committee CEN/TC 434 was initiated to establish a semantic data model of the core elements of an electronic invoice [18]. The information requirement model for the corresponding transactions were aligned with the semantic data model developed by CEN/TC 434, which had an impact on the other IRMs, as all other IRMs are aligned to each other via the business term vocabulary.

In addition to the profiles and transactions, this CWA provides four guidelines as well. These guidelines provide fur-

ther details on how to implement specific use cases with the post-award profiles. The first guideline provides guidance on how to implement a master-data approach using the transactions by CEN/WS BII. The other three guidelines provide guidance on implementing the simplified invoice according to directive 2006/112/EC, on payment initiation and reconciliation, and pre-payments.

VI. Discussion

After presenting the approach and the outcome of CEN/WS BII, the workshop and its work shall be put into context as well as research questions derived that might be of interest for the community of Information Systems (IS).

First of all, the questions regarding the effect of empire building mentioned in the introduction shall be addressed, in particular the question, if the approach and the outcome by CEN/WS BII can serve as a means to reduce the effect of Empire Building.

Referring to Wüster et al. [19], it can be stated that CEN WS/BII moves the break-even point between the costs for standardization and conversion towards conversion. On the one side, the more messages exchange is standardized the more costs will be caused due to missed opportunities by a lack of individuality. For instance, in case of a maximum level of standardization a company may not be able anymore to provide specific services giving the company a competitive advantage.

On the other side, a maximum level of individuality will cause high costs for developing many peer to peer converters between the various formats, costs for the actual conversion of messages, and costs by inappropriate conversions, for instance, loss of information during the process of conversion. As a consequence, there is a trade-off between a maximum level of standardization and a maximum level of individuality. The challenge is to find the “break-even” between these two both extrema, as it is illustrated in Fig. 5.

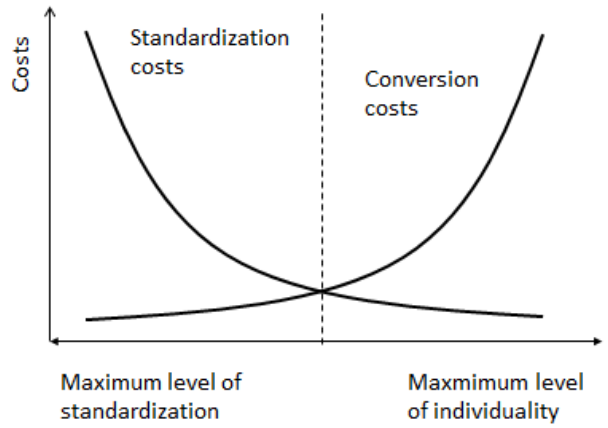


Fig. 5 Trade-off between standardization and conversion [19]

The approach taken by CEN/WS BII moves the point of “break even“ towards the maximum level of individuality, because its reduces the costs for conversion. The IRM with the syntax bindings to UBL messages and UN/CEFACT XML message serve as intermediary language allowing the conversion from one syntax message to another. In the case of the “Item standard identifier”, the syntax binding given in Table 2 states that from the perspective of CEN/WS BII both syntax elements have the same semantic, even if the semantics specified by UBL and UN/CEFACT XML may vary in details. But as long the syntax message is compliant to the corresponding IRM, the syntax elements can be easily converted from one to the other and vice versa.

Furthermore, as there are other standards for syntax messages available, for instance, the German standards BMEcat for exchanging catalogues or ZUGFeRD for exchanging invoices, they can also define a syntax binding for their syntax standard to the corresponding IRM.

This also can be illustrated with the information requirement “Item standard identifier” explained above. There, a syntax binding was defined to a UBL and a UN/CEFACT XML message. By defining a syntax binding to BMEcat as well (cf. Table 4), it now becomes possible to convert the bound syntax elements of UBL, UN/CEFACT XML, and BMEcat into each other.

The semantics of the syntax element INTERNATIONAL_PID defined in the BMEcat specification might be slightly different from the semantics specified for the information requirement “Item standard identifier”. But by using the syntax element with the semantics of “Item standard identifier”, the BMEcat catalogue message becomes compliant with the IRM of CEN/WS BII. This way, the BMEcat community can keep their standard and can address their use cases specific for their community, but allowing a usage of BMEcat compliant with CEN/WS BII. In consequence, the syntax standard BMEcat becomes “interoperable” with the other syntax standards having syntax binding to CEN/WS BII.

Table 4.
Syntax binding for Item standard identifier for UBL, UN/CEFACT XML, and BMEcat

InfReqId	Syntax binding
tir19-092	Catalogue/cac:CatalogueLine/cac:Item/ cac:StandardItemIdentification/cbc:ID
tir19-092	CrossIndustryCatalogue/ CICHSupplyChainTradeTransaction/ IncludedCICLSupplyChainTradeLineItem/ SpecifiedCatalogueTradeProduct/ID
tir19-092	BMECAT/T_NEW_CATALOG/PRODUCT/ PRODUCT_DETAILS/INTERNATIONAL_PID

Consequently, conversions to other relevant syntax standards become easier, as the syntax elements can also be converted to the corresponding syntax elements of UBL and UN/CEFACT XML. This allows communities to define syntax standards for their own special needs as well as being compatible—at least with respect to the core requirements—with other syntax standards in use. The conversions to those standards become less expensive to develop and help to balance heterogeneity and interoperability of information systems in place [20]. In fact, it is argued that communities often just need a core they can adapt to and amend it with their community-specific requirements [21].

Here lies a first field of action, where the IS community can contribute to the standardization work. Although, the ex-

ample might give a good indicator for the benefits of the approach by CEN/WS BII, the question remains, if the approach will keep what it promises, i.e., that small and specialized communities can keep their practice-proved and tested syntax standards on the one side and at the same time are interoperable with other e-procurement systems with acceptable costs. Because, it might be also the case that the approach of CEN/WS BII is just another, but very subtle, means for empire building. It will not squeeze out particular syntax standard out of the market, but rather making all the syntax standards the same by imposing mandatory requirements on them.

Related to this first field of action, there is another field of action related to the second proposition of CEN/WS BII. To make syntax binding to IRMs feasible, these IRMs must cover an accepted and stable set of requirements. As mentioned before, the CWA on methodology and architecture elaborate on the concept of core and core requirements. This part of the CWA gives a definition and requirements on the core as well as hints how to find these core requirements. But this part lacks a precise definition, what a core requirement in practice is, and lacks a sound methodology how to find the core requirements in the various uses cases in the e-procurement chain.

In a wider vision, this requires that the architecture of all types of e-procurement standards used in Europe should be compatible with each other. There are initiatives to promote and ensure this kind of compatibility, such as ISA, the European Commission's program for interoperability solutions for European Public administrations. This initiative developed the so-called European Interoperability Reference Architecture (EIRA). EIRA offers a service-orientated method, models and building blocks to develop, extend, and adapt any kind of e-government solution in Europe in a harmonized manner aiming at achieving interoperability over the whole lifecycle of these systems. In respect to specification by CEN/WS BII the challenge is to harmonize the profile- and process-orientated approach CEN/WS BII with the service-orientated reference architecture of EIRA in a way that the practical needs of users both on the buying and the selling side are taken in account. They have to understand, accept, efficiently use, and incorporate these standards and specifications in their everyday work.

In spite of the benefits of the outcome of CEN/WS BII, there are other limitations. One of them is that possible variants of a product, service, or process are not explicitly modeled. In the terms of CEN/WS BII they are interpreted as extensions or changes of the "Core". But modelling and implementing these extensions or changes may be not easy because of the complexity and major diversities related to the various types of product, service types, and supply chain as well as the worldwide differences procurement regulations on the political, organizational, and technical level. In addition, in some industries and trading areas there exist special order and delivery concepts.

Possible variants may cover product specific differences (liquids, hazards, food), sectors specific differences (health, chemistry, logistics), special supply chain types (projects, automotive supply chains and special order and delivery concepts like vendor managed inventory (VMI) or just-in -time delivery (JIT).

To incorporate these variants in goods, services, processes, and supply chains and to identify the common core requirements as well as provide the means to identify and to specify the core requirements is a third field of action where the IS community can contribute to the field of e-procurement standardization.

VII. Conclusion

In this paper, the work of the third phase of CEN/WS BII was presented and discussed. The outcomes provide a basis for implementing European-wide public e-procurement from notification and tendering to fulfilment and payment. The outcome was discussed arguing that the outcome might reduce the effect of empire building and allows an increased level of individuality by increasing the interoperability of e-procurement systems. Based on this, research questions and themes were outlined that might be addressed by the IS community in the future.

Addressing these research questions, may help to improve the approach of CEN/WS BII. This is important, as CEN has decided to establish the technical committee CEN/TC 440 [22], which picks up the outcome of CEN/WS BII and has the mandate to transform the CWAs into an efficient and acceptable formal European standard for public procurement. In fact, the work done by CEN/WS BII and continued in CEN/TC 440 contributes to one of the "grand challenges of Information Systems research" identified by Becker et al. [23] to "[integrate] information systems in one single virtual space" of e-procurement systems.

Acknowledgment

The authors thank Fred van Blommestein for valuable remarks on earlier drafts of this paper.

References

[1] European Commission, "COM(2012) 179: A Strategy for e-procurement," http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52012DC0179, Retrieved 2016-04-16.

[2] European Commission, "COM(2013) 453: End-to-End E-Procurement to Modernise Public Administration," http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52013DC0453, 2013, Retrieved 2016-02-21.

[3] Single Face To Industry, "Svefaktura BIS 5A 2.0," http://www.sfti.se/standarder/bestallningsprocesssftisvehandel/svefaktura/svefakturabis5a20.2074.html, 2016, Retrieved 2016-02-21.

[4] Direktoratet for forvaltning og IKT, "EHF Invoice and Creditnote 2.0.5," https://vefa.difi.no/ehf/standard/ehf-invoice-and-creditnote-2.0.5/, 2015, Retrieved 2016-02-21.

[5] Forum elektronische Rechnung Deutschland: "ZUGFeRD-Infopakete," http://www.ferd-net.de/front_content.php?idcat=255, 2014, Retrieved 2016-02-21.

[6] F.-D. Dorloff, V. Jahns and V. Schmitz, "E-Business interoperability: A systematization attempt based on the morphology concept," in Electronic Business Interoperability: Concepts, Opportunities and Challenges, E. Kajan, Ed. Hershey: IGI Global, 2011, pp. 1–14. DOI: 10.4018/978-1-60960-485-1.ch001

[7] J. V. Nickerson and M. zur Muehlen, "The ecology of standards processes: Insights from internet standard making," *MIS Quarterly*, vol. 30, pp. 467–488, August 2006.

[8] M. L. Markus, C. W. Steinfield, R. T. Wigand and G. Minton, "Industry-wide information systems standardization as collective action," *MIS Quarterly*, vol. 30, pp. 439–465, August 2006.

[9] J. Backhouse, C. W. Hsu and L. Silva, “Circuits of power in creating de jure standards: Shaping an international information systems security standard,” *MIS Quarterly*, vol. 30, pp. 413–438, August 2006.
[10] V. Fomin and T. Keil, "Standardization: Bridging the gap between economic and social theory", in *Proc. of 21st International Conf. on Information Systems*, Atlanta, 2000, pp. 206–217.
[11] CEN and CENELEC, “Internal regulations part 2: Common rules for standardization work”, http://boss.cen.eu/ref/IR2_E.pdf, 2015, Retrieved 2016-02-21.
[12] CEN/ISSS Business Interoperability Interfaces for Public procurement in Europe (CENBII), http://spec.cenbii.eu/IndexBII1.html, Retrieved 2016-03-05.
[13] Object Management Group, “Business Process Model and Notation 2.0,” http://www.omg.org/spec/BPMN/2.0/, Retrieved 2016-04-23.
[14] CEN Workshop on Business Interoperability Interfaces for public procurement in Europe, http://www.cenbii.eu/, Retrieved 2016-05-09.
[15] European Union, “Directive 2014/23/EU on the award of concession contracts," http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.094.01.0001.01.ENG, Retrieved 2016-05-01.
[16] European Union, “Directive 2014/24/EU on public procurement and repealing Directive 2004/18/EC,” http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.094.01.0065.01.ENG, Retrieved 2016-05-01.
[17] European Union, "Directive 2014/25/EU on procurement by entities operating in the water, energy, transport and postal services sectors and repealing Directive 2004/17/EC,” http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.094.01.0243.01.ENG, Retrieved 2016-05-1.
[18] European Committee for Standardization, “CEN/TC 434 – Electronic Invoicing,” https://standards.cen.eu/dyn/www/f?p=204:7:0::::FSP_ORG_ID:1883209&cs=1E81C9C833655EEDC7010C8D0A2FB786C, Retrieved 2016-05-07.
[19] E. Wüstner, T. Hotzel and P. Buxmann, “Converting business documents: A Classification of problems and solutions using XML/XSLT,” in Proc. of the 4th IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems, Newport Beach, 2002, pp. 54–61.
[20] F.-D. Dorloff and E. Kajan, “Balancing of Heterogeneity and Interoperability in E-Business Networks: The Role of Standards and Protocols”, in *International Journal of E-Business Research*, vol. 8, pp. 15–33, 2012. DOI: 10.4018/jebr.2012100102
[21] T. McGrath, “The Reality of Using Standards for Electronic Business Document Formats,” in *Handbook of Research on E-Business Standards and Protocols: Documents, Data and Advanced Web Technologies*, E. Kajan, F.-D. Dorloff, and I. Bedini, Ed., Hershey: IGI Global, 2012, pp. 21–32. DOI: 10.4018/978-1-4666-0146-8.ch002
[22] European Committee for Standardization, “CEN/TC 440 – Electronic Public Procurement,” https://standards.cen.eu/dyn/www/f?p=204:7:0-::::FSP_ORG_ID:1976650&cs=175E298F320429229DD35C9E22F4E8F76, Retrieved 2016-05-08.
[23] European Commission, “EIRA: European Interoperability Reference Architecture,” http://ec.europa.eu/isa/ready-to-use-solutions/eira_-en.htm, Retrieved 2016-07-08.
[24] J. Becker, J. vom Brocke, M. Heddier and S. Seidel, “In search of Information Systems (Grand) Challenges: A community of inquirers perspective,” in Business & Information Systems Engineering, vol. 57, pp. 377–390, 2015. DOI: 10.1007/s12599-015-0394-0

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 189–193
DOI: 10.15439/2016F466
ACSIS, Vol. 9. ISSN 2300-5963

Developing Coalitions by Pairwise Comparisons: a Preliminary Study

Waldemar W. Koczkodaj
Laurentian University
in Sudbury
935 Ramsey Lk Rd, ON P3E 2C6, Canada
E-mail: wkoczkodaj@cs.laurentian.ca

Anna Tatarczak
Maria Curie-Skłodowska University
in Lublin
Plac Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland
E-mail: anna.tatarczak@poczta.umcs.pl

***Abstract*—In all industries, competition between business organizations is vital to utilize collaborative logistics for planning, forecasting and efficient customer response to optimize the supply chain. Consequently, numerous business organizations build coalitions among themselves making their partnerships more effective.**

The main goal of this study is to investigate how pairwise comparisons can identify near optimal coalition (related to collective intelligence in terms of computer science) for a group of independent business organizations.

Case studies are used to demonstrate the utility of the framework.

***Index Terms*—Collective intelligence, horizontal logistics collaboration, coalition, pairwise comparison, consistency analysis, expert opinion, knowledge management, business process.**

I. Introduction

COLLABORATION mechanism, widely studied in logistics [1], [4], [10], [13], [16], [27], economics [5], [17], [24], [25], [30] and collective intelligence [33], [34], [38], [41], has specified two important issues: alliance formation and gain allocation.

Most models study the process of coalition formation in which all players contemplate forming one big coalition e.g. [37]. However, in many business settings the objective of development cooperation is obtaining the optimal alliance structure. In order to efficiently achieve individual goals and rewards through cooperation, a coalition may not necessary contain all possible collaborations. In realistic scenario, the coalition formation may be required to deal with both possible sub-groups of participants and individual preferences of each potential collaborator. These sub-groups are translated into number of sub-coalitions among independent firms. Finally, these models should reflect the fact that players must take their decision-making problems about in which coalitions they want to be. In particulary, some coalition structure may be more preferable to others. To the best of our knowledge, there is no study that jointly considers these two research problems. For it, we adopt pairwise comparisons (PC) to develop a model of collaboration mechanism.

The paper is structured as follows. Section II is devoted to comprehensive introduction to PC. Section III-A reviews the comprehensive literature concerning the problems of alliance formation in logistics cases. In Section III-B we explain why enterprisers need to collaborate horizontally in logistics, as well as we introduce an alliance formation model among independent companies by PC. Next, in Section IV the method is illustrated by numerical example taken from logistics cases. The last Section concludes the paper, indicating the limits and a list of important directions for further research. Finally, this study is a continuation of what was previously presented at FedCSIS in [18], [6], [2].

II. Preliminaries - pairwise comparisons

Pairwise comparisons may be one of the oldest methods used in science, where we compare entities E_i in pairs when there is no unit of measure in use. In fact, most units are well defined for what we perceive to be "objective" entities. We can effectively measure: distance, weight, temperature, or time. From a mathematical point of view, pairwise comparisons (PC) create a PC matrix (say, M) of values (m_{ij}) of the i-th entity compared with the j-th entity. A small rating scale $[1/c, c]$ is used for i to j comparisons where $c > 1$ is a not-too-large real number (where c is usually 3 to 5 in most practical applications). It is frequently assumed that all the values m_{ij} on the main diagonal are 1 (the case of E_i compared with E_i and that M is *reciprocal*: $m_{ij} = 1/m_{ji}$ for every $i, j = 1, ..., n$, since $x = 1/(1/x)$.

A pioneer of using PC method for elections was Llull in 13th century. Independently, Condorcet [7] used the pairwise comparisons in his publication of 1785 in the context of counting political ballots. In 1860, however, Fechner used this method in [14]. Thurstone [36] described pairwise comparisons method as the Law of Comparative Judgments in [36]. In 1977, Saaty [32] introduced a hierarchy instrumental for practical applications. However, shortcomings of this work have been criticized in [21].

The usefulness of pairwise comparisons approach, as well as reference to Llull were evidenced in one of the flagship ACM publications [12]. Furthermore, Professor Kenneth J. Arrow, the Nobel prize winner, has used "pair" 24 times in his seminal work [3].

We assume that M is a reciprocal PC matrix over R^+, and M is of the form:

$$M = \begin{bmatrix} 1 & m_{12} & \cdots & m_{1n} \\ \frac{1}{m_{12}} & 1 & \cdots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{m_{1n}} & \frac{1}{m_{2n}} & \cdots & 1 \end{bmatrix}$$

III. Coalition Formation in Horizontal Logistics Collaboration

A. Literature review

In the past few decades, great attention has been placed on the logistics collaborations, which is concerned to be attained when two or more business organizations enter into a partnership for the purpose of improving the goals of supply chain in an effective and efficient way. Logistics collaborations are categorized as vertical and horizontal.

In this paper, we focus our attention on horizontal cooperation in logistics. It is defined as collaborations between companies that occurs on the same level of the supply chain [10] (e.g. collaboration among suppliers). The primary objective of the horizontal cooperation for each members is reduction of the costs. At the same time, coalition members can also achieve efficiency improvement and better customer service. Based on the research by Cruijssen [9], motives for horizontal cooperation can be categorized as listed in Table I.

The theory of coalition formation have traditionally been analyzed based on cooperative game theory. This popular approach was rewarded when L.S. Shapley received the Nobel Memorial Price in Economics Sciences 2012 together with A. E. Roth. Different cases of coalition formation and cost allocation methods are reported in the recent literature. However, the most popular are five types: transportation planning, traveling salesman problem, vehicle routing, joint distribution, and inventory related problems.

The horizontal collaboration in logistics is profitable, but the categories of impediments were also defined. They are: partner selection, negotiation and coordination, information and communication technology, determining and dividing the gains [9]. Taken into account the main barriers to collaboration, which are the problems of partner selection and communication technology factors we proposed a heuristic model on coalition determination in Section III-B through the application of PC.

Given this background, coalition formation among independent firms has received a considerable amount of attention in recent research, and has proven to be adaptable in most practical cases.

B. Model of study

Game theory [26] provides a good notion for several concepts which are useful in coalition formation e.g. players, coalition, stability, super-additivity. Let N be a finite set of *players* and X a finite set of *states.* For convenience, the players are numbered such that $N = \{1, 2, ..., n\}$. A *coalition* is a nonempty subset S of N. If two players do not cooperate, they belong to different coalitions. The *grand coalition* refers to $S = N$.

For the further discussion, we assume that the set of players is finite equal $N = 3$. Particulary, this study examines the subject of alliance formation among n $(n \geq 3)$ independent firms that have a great potential to create synergies. In particular, we consider a group of three business organizations: X_1, X_2, X_3. In order to save on transportation costs, business organizations want to collaborate with each other, but usually, players do not know in which combination. Some strategic alliances and alliance partners can be more preferable and more probable than others. For $n = 3$ there are four possible situation (which are non-empty and non-singletons), namely

- X_1 can create a coalition with X_2,
- X_1 can create a coalition with X_3,
- X_2 can create a coalition with X_3,
- grand coalition X_1, X_2, X_3 could be formed.

When all business organizations create a grand coalition, it is the most concerning alliance formation problem when the collaboration is super-additive. It is addressed in [28], [29], [23], [22], [15], [11]. The *super-additivity* means that whatever two separate coalitions creating union get at least as much as they get independently. Under these circumstances, it may be expected to form a grand coalition. Contrary to many complex logistics issues (as it is in the case of our model), super-additivity is too exigent.

Through coalition formation process, the cost reduction could be achieved by applying pooling strategies on logistics planning and optimization, which is defined as common usage logistics resources e.g. vehicles, platform, software tool. Generally, in the logistics pooling investigation, we divide pooling collaborations into four categories (listed in Table II) with respect to different collaboration preferences and coordination cost e.g. communication costs, investment in IT [40]. These factors impact the potential sub-coalition decision-making scheme and thus should be considered in modeling process. As coordination costs are negligible groups grows by joining new members and the corresponding cooperative game is super-additive. As coordination costs increase, the games becomes non-super-additive.

From now on, we focus on situation when independent business organizations arrange pooling collaborations among them to minimize their individual costs with individual optimum preferences (C_4 case in Table II). In this case, a collaboration scheme based on sub-coalition (with sub-groups of participants) seems to be optimal and preferable. A natural question arises: "Which coalitions can be expected to be formed?" Our objective is to give an answer to this question. By applying PC, we decide which sub-coalition will be optimal for the alliances of company. The following three steps have been used.

Step 1. **Completion of the pairwise comparisons matrix.** To achieve an efficient collaboration by our model, the first step is to select all possible subsets of the members in a group of companies. For n members it would lead us to $2^n - n - 1$ sub-coalition, which are

TABLE I
MOTIVES FOR HORIZONTAL COOPERATION IN LOGISTICS CONTEXT [9].

Cost and productivity	**Customer service**	**Market position**	**Other**
- Cost reduction - Learning and internalization of tacit, collective and embedded knowledge - More skills labour force	- Complementary goods and services - Ability to comply to strict customer requirements/improved service - Specialization	- Penetrating new markets - New products development - Serving larger firms - Protecting market share - Faster speed to market	- Developing technical standards - Accessing superior technology - Overcoming legal/regulatory barriers - Enhancing public image

TABLE II
SELECTED CATEGORIES OF POOLING CASES [40].

Coordination cost	**Global optimum**	**Individual optimum**
Negligible	C_1 : super-additive collaboration with global optimum preferences	C_2 : super-additive collaboration with individual optimum preferences
Significant	C_3 : non-super-additive collaboration with global optimum preferences	C_4 : non-super-additive collaboration with individual optimum preferences

non-empty and non-singletons. Next, each company evaluates two criteria connected with possible structure of coalition in term of their relative importance. Index value from 1 to 4 (and its inverse $1/4$ to 1) are used and entered row by row into a cross-matrix. All gradations are possible in between. Each pairwise comparisons $a_{ij} \in [1/4, 4]$ represents the scaled relative importance scores of element i as compared to element j. In practice, values below the main diagonal do not need to be entered, since they are reciprocal to the corresponding values in the upper triangle. As a result, we get a pairwise comparisons matrix of size $m \times m$, where m denotes the number of items to compare. Formally, for n business organization $m = 2^{n-1} - 1$.

Step 2. **Calculating the criteria weight.** Once the firms provide all the $m(m-1)/2$ comparisons, the goal is to find a positive weight vector $w = [w_1, ...w_m] \in R^m$ such that the pairwise ratio of the weights, w_i/w_j are as close as possible to the matrix elements $a_{ij},\ 1 \leq i, j \leq m$. For this purpose, we use the geometric means (GM) method.

Step 3. **Collaboration decision.** In this step, collaborators try to establish the structure of the coalition. This coalition for which the mean value is highest, is the most preferable for all potential partners and as a result successful collaborations groups should be established. However, in practice, there are two possibilities for this step: an agreement is achieved (collaboration relationship is established) or individual members of group deviate, the others move back to Step 1.

By constructing a PC matrix, we are able to identify the most desirable coalition for business organizations with individual preferences expressing incentives to collaborate. Following this collaboration process and specifying the technical details, this study provides a solution to horizontally cooperating companies who are reluctant to share sensitive information and want to avoid making excessive communications.

IV. RESULTS AND ANALYSIS

In this Section, we use a numerical example to apply presented PC method to determined alliance formation among independent business organizations, which have significance motivation to cooperate. This motivation has been frequently associated with the following drivers of horizontal collaboration: cost reduction, service improvement, market position, skill and knowledge sharing, investment and risk sharing, emission reduction, congestion reduction [40] as well as protect environment and mitigate climate change [35].

Let suppose we have a group of three companies (labeled X_1, X_2, X_3), then we separate candidate into possible collaboration groups (which are non-empty and non-singleton), namely:

$$\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}, \{X_1, X_2, X_3\}$$

In our case, each company based on their individual satisfaction (connected with drivers of horizontal collaboration) has to fill one pairwise comparisons matrix. Every business organization needed to determine 3 ratios: m_{12}, m_{13}, m_{23}, which are the part of following matrix:

$$M = \begin{bmatrix} 1 & m_{12} & m_{13} \\ \frac{1}{m_{12}} & 1 & m_{23} \\ \frac{1}{m_{13}} & \frac{1}{m_{23}} & 1 \end{bmatrix}$$

This step is the most complicated and should consist of some technical tools to design efficient method for filling the matrix. In each matrix on the main diagonal, we have number 1 due to the fact that it represents a relative ratio of a criterion against itself. All three PC matrixes are obtained and presented in Table III.

After each business organization sets its preference of alliance with other partners, the PC matrix is constructed and weights are computed as the geometric means of rows;

TABLE III
PAIRWISE COMPARISONS MATRIX FOR EACH COMPANY.

Company X_1 preferences			
	$\{X_1, X_2\}$	$\{X_1, X_3\}$	$\{X_1, X_2, X_3\}$
$\{X_1, X_2\}$	1	1,5	3
$\{X_1, X_3\}$	0,67	1	3,6
$\{X_1, X_2, X_3\}$	0,33	0,28	1
Company X_2 preferences			
	$\{X_1, X_2\}$	$\{X_2, X_3\}$	$\{X_1, X_2, X_3\}$
$\{X_1, X_2\}$	1	2	2
$\{X_2, X_3\}$	0,5	1	2
$\{X_1, X_2, X_3\}$	0,5	0,5	1
Company X_3 preferences			
	$\{X_1, X_3\}$	$\{X_2, X_3\}$	$\{X_1, X_2, X_3\}$
$\{X_1, X_2\}$	1	2,5	1,3
$\{X_2, X_3\}$	0,4	1	4
$\{X_1, X_2, X_3\}$	0,77	0,25	1

the weights of their coalitions structure were calculated using Concluder [8].

TABLE IV
WEIGHT EVALUATION BASED ON PAIRWISE COMPARISONS.

	$\{X_1, X_2\}$	$\{X_1, X_3\}$	$\{X_2, X_3\}$	$\{X_1, X_2, X_3\}$
X_1	0,48	0,39	0	0,13
X_2	0,49	0	0,31	0,2
X_3	0	0,46	0,36	0,18
Average	0,32	0,28	0,22	0,17

In each row, there is one 0 value. It refers to the situation when a company does not have incentive to create a coalition with itself. Next, based on the results, a suitable coalition structure is constructed. It is supported by PC as a guidance tool for building coalitions. In the represented case, the highest preferences have been assigned to $\{X_1, X_2\}$, while on the second place we have $\{X_1, X_3\}$ coalition, and the least profitable sub-coalition is a grand coalition $\{X_1, X_2, X_3\}$.

V. NUMBER OF COMPARISONS

The case study indicates that our model is attractive for smaller numbers of companies from 3 up 5. From a practical point of view, the number of comparisons that should be made by one company is easy to handle. For bigger number of potential partners ($n \geq 6$), the comparisons are increasing exponentially. In Table V we see that for 7 companies, number of comparisons that should be made by one partner is 1953.

Table V shows that exponential growth does not cause comparisons increases for up 5 business organizations. Generally, when we have n company, the number of needed comparisons are presented in Table VI.

VI. CONCLUSIONS AND FUTURE RESEARCH

This study has demonstrated that using pairwise comparisons, as a framework to model cooperation in logistics, can produce valuable results. We have demonstrated in this study that pairwise comparisons can be effectively used for formation of coalitions when the coalition forming solutions follow the *satisficing principle* introduced by Herbert A. Simon in 1956 and well described in [39].

What is more important that this study initiates a new direction of research in important logistics problems. One of them is the open problem of the bargaining power of collaborative members in business negotiations. In current logistics practice, we observe that a person, group, or organization has ability to enforce their preferences. Such bargaining power is closely related to influence, financial and organizational power, size or status of the collaborators. The bargaining power directly influences the potential group decision. It should be considered in the alliance formation process.

Another research direction worth pursuing on coalition formation by PC is the characterization of the complementary coalition. The complementary coalition is the coalition of the company that is not included in the primary coalition, e.g. we have 5 business organizations that want to use the proposed method. By it, we find that coalition X_2, X_5 is near optimal coalition structure. What would be now the complementary coalition from the set X_1, X_3, X_4? The third research direction is to investigate the coalition structure with some additional constraints. For instance, the number of coalition members in a coalition could be limited to $k,\ 2 \leq k \leq n$. The question now is how to form and search for such coalition structure. These issues are subject for our future research.

ACKNOWLEDGMENT

The authors are grateful to Grant O. Duncan (Team Lead, Business Intelligence and Software Integration, Health Sciences North, Sudbury, Ontario) and Tyler Jessup, Laurentian University, for their help with proofreading this text and development of the software system based on the presented theory.

REFERENCES

[1] Agarwal, R.; Ergun, Ö., Network design and allocation mechanisms for carriere alliances in liner shipping, Operations Research, 58(6); 1726 - 1742, 2010.
[2] Alqarni, M.; Y. Arabi, Y.; Kakiashvili, T.; Khedr, M.; Koczkodaj, W.W.; Leszek, J.; Przelaskowski, A.; Rutkowski, K., Improving the predictability of ICU illness severity scales, Federated Conference on Computer Science and Information Systems (FedCSIS), Szczecin, IEEE Conference Publications, 11 - 17, 2011.
[3] Arrow, K.J., A Difficulty in the Concept of Social Welfare, Journal of Political Economy, 58(4): 328-34, 1950.
[4] Audy, J.F.; D'Amours, S.; Rönnqvist, M., An empirical study on coalition formation and cost/savings allocation, International Journal of Production Economics, 136(1): 13 - 27, 2012.
[5] Aumann, R.J.; Myerson, R. B., Endogenous formation of links between players of coalitions: an application of Shapley value, University Press, Cambridge, 175-191, 1988.
[6] Babiy, V.; Janicki, R.; Wassyng, A.; Bogobowicz, A.D.; Koczkodaj, W.W., Selecting the best strategy in a software certification process, Federated Conference on Computer Science and Information Systems (FedCSIS): 53 - 58, 2010.
[7] Condorcet, M., The Essay on the Application of Analysis to the Probability of Majority Decisions, Paris: Imprimerie Royale, 1785.
[8] Concluder, $https\ :\ //sourceforge.net/projects/concluder/$, accessed 2016-05-01.

TABLE V
NUMBER OF COMPARISONS.

Number of company	Possible sub-coalition	Number of PCM	Dimension of PCM	Comparisons in one PCM	Total number of comparisons
3	4	3	3	3	9
4	11	4	7	21	84
5	26	5	15	105	525
6	57	6	31	465	2 790
7	120	7	63	1 953	13 671
8	247	8	127	8 001	64 008
9	502	9	255	32 385	291 465
10	1 013	10	511	130 305	1 303 050

TABLE VI
FORMULAS FOR NUMBER OF COMPARISONS.

n	- number of companies,
$2^n - n - 1$	- number of all possible sub-coalitions,
$2^{n-1} - 1$	- dimension of one PCM,
$(2^{n-1} - 1)(2^{n-2} - 1)$	- number of comparisons in one PCM,
$n(2^{n-1} - 1)(2^{n-2} - 1)$	- total number of comparisons,

[9] Cruijssen, F., Horizontal Cooperation in Transport and Logistics, Dissertation thesis, University of Tilburg, 2007.

[10] Cruijssen, F.; Dullaert, W.; Fleuren, H., Horizontal cooperation in transport and logistics: A literature review, Transportation Journal, 207 (3): 22 - 39, 2007.

[11] Dai, B.; Chen, H., A multi-agent and auction-based framework and approch for carrier collaboration, Logistics Research, 3: 101 - 120, 2011.

[12] Faliszewski, P.; Hemaspaandra, E.; Hemaspaandra, L. A., Using Complexity to Protect Elections, Communications of the ACM, 53(11): 74-82, 2010.

[13] Fang, X.; Cho, S.-H., Stability and endogenous formation of inventory transshipment networks, Operations Research, 62(6): 1316 - 1334, 2014.

[14] Fechner, G., Elemente der Psychophysik, 1860.

[15] Frisk, M.; Göthe-Lundgren, M.; Jörnsten, K.; Rönnqvist, M., Cost allocation in collaborative forest transportation, European Journal of Operational Research, 205: 448 - 458, 2010.

[16] Guajardo, M.; Rönnqvist, M, Operations research models for coalition structure in collaborative logistics, European Journal of Operational Research, 240(1), 147 - 159, 2015.

[17] Kahan, J.P.; Rapoport, A., Theories of coalition formation, Lawrence Erlbaum Associates, 1984.

[18] Kakiashvili, K.; Koczkodaj, W.W.; Phyllis Montgomery, P.; Passi, K.; Tadeusiewicz, R., Assessing the Properties of the World Health OrganizationŠs Quality of Life Index, Federated Conference on Computer Science and Information Systems (FedCSIS), Wisla, IEEE Conference Publications, 151 - 154, 2008.

[19] Kefi, M.; Ghedire, K., A multi-agent model for the Vehicle Routing Problem with Time Windows, WIT Transactions on The Built Environment, 75, 2004.

[20] Kemahlioglu-Ziya, E.; Bartholdi III, J.J., Centralizing inventory in supply chain by using Shapley value to allocate the profits, Manufacturing & Service Operations Management, 13(2), 146-162, 2011.

[21] Koczkodaj, W.W., Mikhailov, L., Redlarski, G., Soltys, M., Szybowski, J., Tamazian, G., Wajch, E., Yuen, K.K.F., Important Facts and Observations about Pairwise Comparisons, Fundamenta Informaticae, 144: 1-17, (2016)

[22] Krajewska, M.; Kopfer, H.; Laporte, G.; Ropke, S.; Zaccour, G., Horizontal cooperation among freight carriers: request allocation and profit sharing, Journal of the Operational Research Society, 59: 1483 - 1491, 2008.

[23] Lozano, S.; Moreno, P.; Adenso-Díaz, B.; Algabaí, E., Cooperative game theory approach to allocating benefits of horizontal cooperation, European Journal of Operational Research, 229: 444 - 452, 2013.

[24] Luce, R.D.; Raiffa, H., Games decisions, John Wiley & Sons, New York, 1957.

[25] Myerson, R.B., Game theory: analysis of conflict, Harvard Unversity Press, 1991.

[26] G. Owen, Game theory, London, UK: Academic Press, Oct. 1995.

[27] Özener, O., Developing a collaborative planning framework for sustainable transportation, Mathematical Problems in Engineering, art. ID 107102, 2014.

[28] Özener, O.; Ergun, Ö., Allocationg costs in a collaborative transportation procurement network, Transportation Science, 42: 146 - 165, 2008.

[29] Perea, F.; Puerto, J.; Fernández, F.; Modeling cooperative on a class of distribution problems, European Journal of Operational Research, 198: 726 - 733, 2009.

[30] Rapoport, A.; Kahan, J.P.; Funk, S.G.; Horowitz, A.D., Coalition formation by sophisticated players, Springer, 1979.

[31] Saad, W.; Han, Z.; Basar, T.; Debbah, M.; Hjorungnes, A., Hedonic coalition formation for distributed task allocation among wireless agents, IEEE Transactions on Mobile Computing, 10(9), 1327 - 1344, 2011.

[32] Saaty, T.L., A scaling method for priorities in hierarchical structures, Journal of Mathematical Psychology, 15: 234-281, 1977.

[33] Sandholm, T.; Larson, K.; Anderson, M.; Shehory, O.; Tohme, F., Antytime coalition structure generation with worst case guarantees, Proceedings of the Fifteenth National Conference on Artifical Intelligence, 46 - 53, 1998.

[34] Sandholm, T.; Lesser, V.R., Coalitions among computationally bounded agents, Artifical Intelligence, 94: 99 - 137, 1997.

[35] Simchi-Levi, D.; Simchi-Levi, E.; Kaminsky, P., Designing and managing the supply chain: concepts, strategies, and cases, New-York: McGraw-Hill, 1999.

[36] Thurstone, L.L., A law of comparative judgement, Psychological Review, 34: 278-286, 1927.

[37] Ulrike, L. - W., Pauoff divisions on coalition formation in a three-person characteristic function experiment, Journal of Economic Behavior & Organization, 17(1): 183 - 193, 1992.

[38] Wi, H.; Oh, S.; Mun, J.; Jung, M., A team formation model based on knowledge and collaboration, Expert Systems with Applications, 36(5): 9121 - 9124, 2009.

[39] Winter, S.G., The satisficing principle in capability learning, Startegic Management Journal, 21(10 - 11): 981 - 996, 2000.

[40] Xu, X., Collaboration mechanism in the horizontal logistics collaboration, Dissertation thesis, 2014.

[41] Zlotkin, G.; Rosenschein, J.S., Coalition, cryptography, and stability: Mechanism for coalition formation in task oriented domains, AAAI, 432: 87 - 94, 1994.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 195–202
DOI: 10.15439/2016F209
ACSIS, Vol. 9. ISSN 2300-5963

The model of delivering an IT product designed to activate and support senior citizens in Poland

Wieslawa Gryncewicz, Robert Kutera, Maja Leszczynska, Beata Butryn
Wroclaw University of Economics, ul. Komandorska 118-120, 53-345 Wroclaw, Poland
Email: {wieslawa.gryncewicz, robert.kutera, maja.leszczynska, beata.butryn}@ue.wroc.pl}

Abstract— **In the article the authors would like to find the most suitable model of delivering an IT product for elderly people designed to activate them and support their everyday activities. The article consists of a few parts. At the beginning the situation and forecasts for the population of elderly people in Poland are presented. Later the methodology of conducted research is described. In the next part the analysis of the factors influencing the activation and support of elderly people has been made. On the basis of the results of the PEST analysis key roles of stakeholders of the micro- and macroenvironment of the particular IT product are identified and their influence on the delivery model is explained. The core functions of each role are elicited and briefly described in the context of the IT product delivery model for the chosen target group.**

I. Introduction

WHAT can be observed in many European societies is a demographic shift which causes significant changes in the proportion between the young and old population. The situation in Poland is very similar, however Poland, in contrast to many European countries, still seems to be a demographically young country. However, the projections for the future show a slight decrease in the total population of Polish citizens, in 2050 it will decrease by 4,5m, while the amount and percentage share of elderly people will increase from 21.5% (8,3m) in 2013 to 40,4% (13,7m) in 2050 [1]. This incorporates significant challenges for the Polish authorities and society itself. It is an impending threat to the economic and social stability in the country because of a great financial strain on the retirement system. As a result, the economic situation of the elderly population will get worse and they might be pushed into social exclusion.

Therefore, different parties undertake various activities to prepare for this situation. European Union funds are allocated among different R&D programs aimed at the improvement of the situation of elderly people like "Active and Assisted Living Research and Development Programme", an initiative for fostering the emergence of innovative ICT products and services, allowing seniors to live independently, in order to improve their quality of life and autonomy and reduce the cost of their care [2]. Among the Horizon 2020 programs there is also the Work Programme 2016-2017 for Societal Challenge 1 (SC1 - related to Health Demographic Change and Wellbeing) of the Horizon 2020 programme to complement, support and add value to the policies of the Member States aimed at improving the health of the EU citizens and reduce health inequalities by promoting health, encouraging innovation in health, increasing the sustainability of health systems and protecting the Union citizens from serious cross-border health threats [3]. The central government is trying to adjust the retirement system and it also organizes some funding programs, like "Government Program for Social Activity of Elderly People for years 2014-2020", whose aim is to improve the quality and level of life of elderly people in order to provide them with the possibility to age in a dignified manner through social activity [4]. There are also other initiatives organized by NGOs or private investors, like "Seniors in Action" aimed at enabling the realization of social projects by elderly people. There are many NGOs that support elderly people in their lives, some of them have national coverage, others – local. Local authorities try to support elderly people as well. They organize dedicated units to deal with matters of the elderly people, which often cooperate with NGOs, hospitals and nursing homes.

All these initiatives aimed at supporting elderly people can be supported by innovative information and communication technologies. Flexible web programming technologies, mobile devices (smartphones, tablets, wearables) with different kind of sensors (incl. health sensors) and effective wireless transmission technologies can significantly make life easier for elderly people if an IT product is customized for their needs.

At the European market, as well as at the Polish one, there are many web portals through which you can arrange a visit to a doctor or find a plumber (proreferral.com, thumbtack.com, myhammer.de, hassle.com, skilldigger.com, getdoido.com, sirlocal.pl, zamow-fachowca.pl, favore.pl, znanylekarz.pl, freelancer.pl). However, they are not dedicated to elderly people and therefore are not suited to their needs. It is important to offer seniors a solution that:

- supports them in many dimensions (providing services of different kind within a common platform),

- stimulates their everyday activity,
- ensures proper interaction with elderly people through the interface adapted to their perceptions and needs.

The lack of such a technical solution is noticed also by different units and organizations which deal with considered target group.

Therefore the main aim of the article is to find the most suitable model of delivering an IT product for elderly people in order to activate and support them. For this purpose an analysis of the factors influencing the activation and support of elderly people has been made. It is aimed at finding key roles of stakeholders of the micro- and macroenvironment of the particular IT product as well as their core functions in the context of the IT product delivery model for the chosen target group.

II. METHODOLOGY

The research on activation and support of elderly people in Poland with an IT product was preceded by a wide review of scientific and professional literature, as well as of the market reports with rich statistical data. After the identification of a knowledge domain the analysis of the gathered information was made. PEST Analysis was a method chosen for that purpose. It is a tool designed to analyse macro environmental factors [5]. The factors are classified into four different categories which cover political, economic, technological and social factors. As a result, there could be defined an environmental context of the issues analysed and the directions of micro environmental changes. Such an analysis is most frequently used to specify essential environmental sectors which influence a particular organization and its operating strategy [6] [7].

The purpose of the analysis resulting from a thorough understanding of the socio-political, economic and technological context was to identify the key characteristics of the IT product designed to activate and support this social group. They covered such factors as: the living standards of elderly people in Poland, their digital competencies and the infrastructural conditions.

The application of PEST tool resulted in the preliminary verification of the Polish market potential for the IT product dedicated to elderly people. It also allows to define the general profile of an IT product which may, after its adjustment to specific macroeconomic conditions, respond to the needs and expectations of the prospective recipient. The examination procedure involved the following steps:

1) Specifying through brainstorming the most significant factors to be taken into account in PEST Analysis, namely political, economic, social and technological factors.
2) Verification of the available research reports, including the statistical ones, and of the available resources in order to carry out a detailed analysis of the factors and in order to specify their impact and likelihood.
3) Specifying the influence of the factors on the IT product profile by defining the characteristics thereof.

The results of the analysis were later utilized for defining the model of the IT product delivery. For the graphical representation an onion model was chosen. It is a graph-based diagram template, which can be useful for understanding the interrelationships between an IT product and its stakeholders from micro- and macroenvironment [8][9]. Its structure can reflect the complexity of the model and strength of impact of particular roles of stakeholders by placing them on the proper layer. The first step in building the model was to develop a list of the potential roles of the stakeholders. Then the roles were assigned to the proper layers of the model: micro- or macroenvironment of the IT product. At the end, the core functions of each role were defined, analyzed and described.

III. PEST ANALYSIS – CONCLUSIONS AND REMARKS FOR AN IT PRODUCT

The Authors evaluated the factors which had been selected at the first stage of PEST Analysis by specifying their impact (on a scale from -2 to 2, where -2 was means factors with a very negative impact, +2 means factors with a very positive impact and 0 means factors of a neutral character) and likelihood (on a scale from 0 to 1, where 0 means unlikely phenomena and 1 means phenomena certain to happen). The influence was evaluated by multiplying one of the aforementioned factors by another. The detailed description of the whole studies and their results were presented during EHST conference last year [10].

Furthermore, the Authors determined in discussion the characteristics of the IT product dedicated to elderly people. The said characteristics constitute a response to a given factor and if the factor is negative, they constitute an antidote which is capable of eliminating its influence. At this stage, it is also possible to notice that according to the Authors, what influences the IT product dedicated to elderly people in the strongest, positive way is a high level of informatisation in Poland and what has the most negative influence are the biological, psychological and social barriers in the IT perception related to aging. The summary of the PEST Analysis effects is presented in Table 1.

On the basis of PEST Analysis, the Authors observed that what has the strongest positive influence on such a product is a high level of informatisation in Poland. Due to the existence of a developed infrastructure and due to the falling costs of its use, elderly people have a better access to modern IT products and to the Internet. Furthermore, both the domestic and European policies support initiatives dedicated to the analysed social group by providing the source of financing. Thus, IT companies have a possibility to provide senior citizens with a free access to their products and services. At the same time, the Software-as-a-service (SaaS) model is gaining importance as a form of software delivery. In that model the entire infrastructure along with the software is under the control of the service provider, while the user retains control over his or her data [11].

TABLE 1 PEST ANALYSIS RESULTS

Factor	Impact (from -2 to 2)	Likelihood (from 0 to 1)	Influence (impact x likelihood)	IT product expected characteristics
Political factors				
European and domestic policies facilitating activation of senior citizens	+2	0,5	1	Free
Domestic legal framework	-1	1	-1	Compliant with the legal framework
Prolonging working life	1,8	0,6	1,08	Oriented at offering services and entering into transactions
Economic factors				
Structure of income/expenses	-1	0,7	-0,7	Free
Level of wealth	1,2	0,5	0,6	Free
Social factors				
Age-related biological, mental, social barriers in IT perception	-1,6	0,8	-1,28	Adjusted to elderly people's perception Help desk support provided
Social mobility	1,4	0,3	0,42	Community oriented (relations and communication) Mobile Integrated with popular messengers
Technological factors				
Informatisation level in Poland	2	1	2	Available online
Level of acceptance of technology by citizens	-1,8	0,3	-0,54	Help desk support provided
Condition of telecommunication market in Poland	0,6	0,9	0,54	Using popular communication channels (text messages, e-mail)
Software provision method	1,5	0,7	1,05	Available in SaaS model
Easiness of software developing	0,4	0,9	0,36	Using web standards Open to integration
Technological progress	0,6	0,7	0,42	Easily expandable

At the same time negative influences of several factors of the analysed environment were observed. While creating the profile of the IT product dedicated to elderly people one should take into consideration biological, psychological, social and legal barriers which constitute an obstacle for the users in question. Therefore, it is important to create suitable IT products which will respond to the needs and perception of elderly people and to undertake measures designed to educate them in this field. This will allow senior citizens to benefit from their intellectual capital, experiences and skills. Furthermore, this will help to eliminate the generation gap as well as the digital exclusion of elderly people.

IV. The Identification of the Core Stakeholders' Roles and Their Functions in the IT Product Delivery Model

What was identified at the first stage of research on the IT product delivery model dedicated to elderly people was the role which might be performed by the stakeholders functioning in the product's environment. The aforementioned roles are graphically presented in Fig. 1. It should be emphasised that the presented model does not exclude either the situation in which different roles are performed by different stakeholders or the situation in which one stakeholder performs distinct roles.

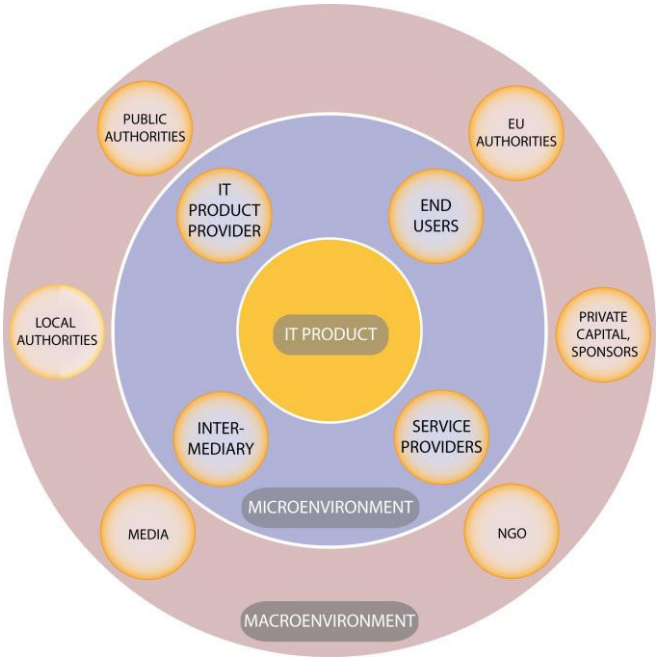


Fig. 1. Key roles of the stakeholders of IT product in the context of its delivery model for elderly people

Furthermore, the model demonstrates the roles of stakeholders from microenvironment, which directly interact with the product. Additionally, it presents the roles of stakeholders from macroenvironment, which interact with

the product in an indirect manner and which create the context and environment where both the product and the stakeholders from microenvironment function. In order to present the IT product delivery model dedicated to elderly people the Authors used the Onion Model, whose layers allowed them to clearly demonstrate and analyse the elements of the micro- and macroenvironment [10].

The roles of stakeholders are the following:

- End users - senior citizens,
- Service providers - entities offering services and products to senior citizens,
- IT product provider - the entity responsible for the IT product development, its delivery to the other stakeholders and technological support,
- Intermediary - the entity which mediates between senior citizens, service providers, and the IT product provider.

The roles of stakeholders from macroenvironment are the following:

- EU authorities,
- Public authorities,
- Local authorities,
- Private capital, sponsors,
- Public benefit organisations – NGO,
- Media.

As it has already been indicated above, the stakeholders from macroenvironment are those which create the environment where the other elements of the model can function. The said environment can be understood as the legal regulations which shape not only the social and economic conditions but also available ways of financing innovative solutions for senior citizens (private and public capital), bringing senior citizens together and activating them.

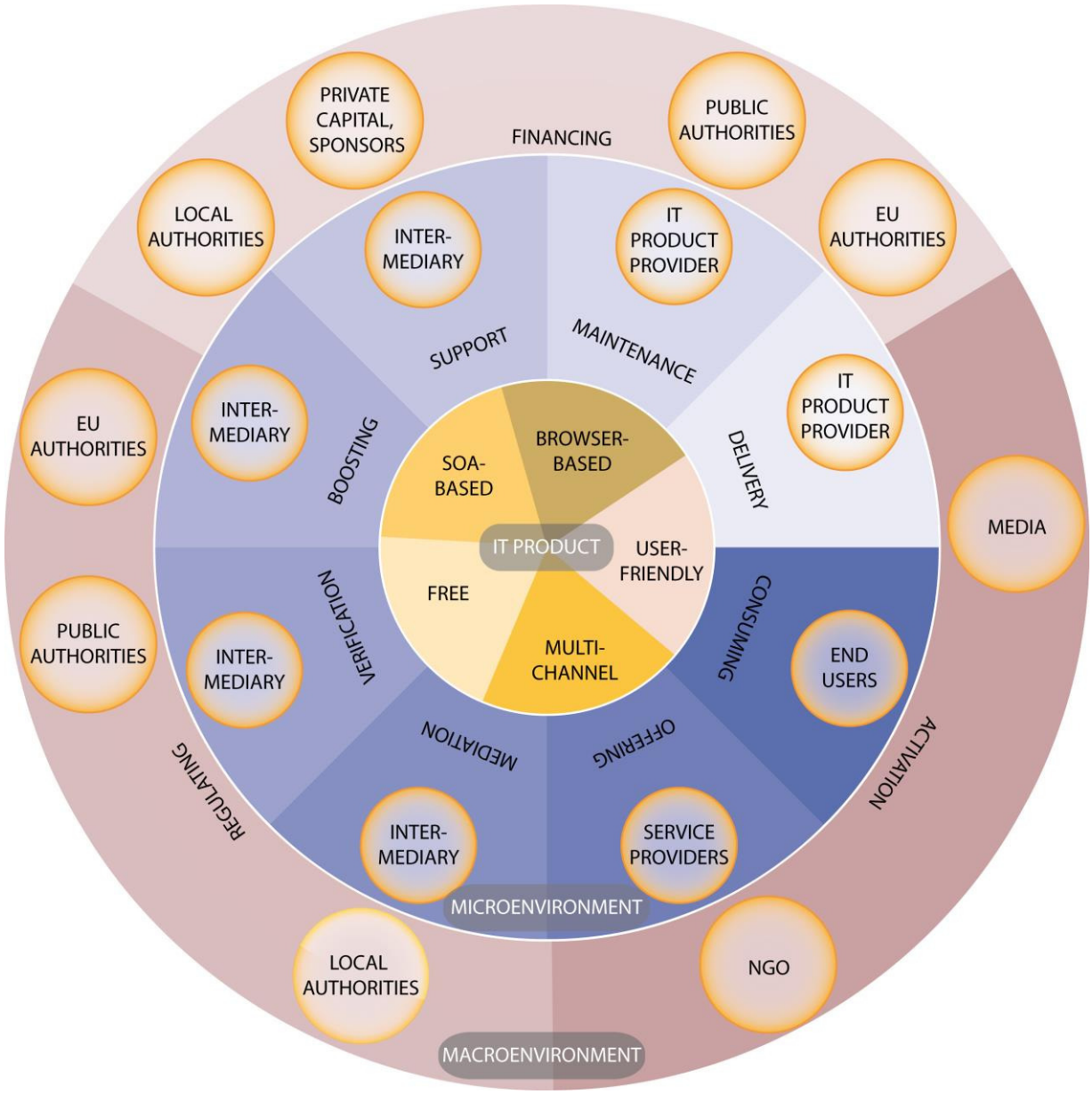


Fig. 2. IT product delivery model for activating and supporting elderly people

The next step in the research on the presented model was to define the characteristics of the IT product itself as well as to indicate the functions performed by particular roles located in the respective layers of the model. Owing to this, it was possible to present the model of a comprehensive character, which ultimately covered the product characteristics, roles performed by particular stakeholders and their functions. This has been illustrated in Fig. 2. It should be noted that the presented model of the IT product delivery has been developed on the basis of the PEST analysis results. Its elements have been defined in such a way that they constitute a response to the factors selected during the analysis and if the factor is negative, they constitute an antidote which is capable of eliminating its influence.

Among the characteristics defining the shape of the IT product the Authors indicated the following:

- The product being available to senior citizens for free,
- The product being available in the SOA model,
- The product being available through the browser without the necessity to install it,
- The interface being unified for all the offered services and featuring improvements dedicated to senior citizens,
- Multichannel - the IT product ability to work with different communication channels.

The fact that the product should be made available to senior citizens for free responds to this factor selected in PEST analysis which indicates that elderly people in Poland have low incomes and therefore they will refuse to pay for the usage of an application of any kind. Another factor selected for the purpose of PEST analysis which has an influence on the IT product characteristics is the one related to the low level of digital competences among senior citizens. What has been suggested in order to eliminate this factor is the solution in which it will be possible to use the IT product without the necessity to install it. The aforementioned solution will be available due to the fact that the product will be built in compliance with the SOA (Service-Oriented Architecture) principleswhat ultimately will make it available as a service from the browser level. In practice, this allows to access the functionalities of the product and its interfaces just after entering the proper web page address into a web browser. The unified interface featuring improvements dedicated to senior citizens constitutes, in turn, the response to the barriers related to the limited perception of IT solutions, which results from biological and mental aging. Multichannel, which provides an access to different communication channels, shall enable efficient communication between both senior citizens and service providers and among senior citizens themselves. It is worthwhile to mention that the channels should be selected according to the criterion of usefulness for the target group. PEST analysis indicates that among the web technologies, senior citizens most frequently choose voice messengers (e.g. skype).

The next stage of building the IT product delivery model designed to support senior citizens' independence in life consisted in defining the functions which are performed by the particular roles. They shall be characterised below.

The term "end users" should be understood in this model as senior citizens who use the IT product to order those services which will allow them to live self-reliant, independent lives. Therefore, they perform the "consuming" function in this model. In this function they register and then log into the IT product, thus gaining the access to their profile, to service providers' offers and to the mechanisms of ordering, appointment coordination and service quality evaluation. The interface available is unified and its appearance has been tailored to the needs of senior citizens. It becomes their single point of contact and allows them not only to order and evaluate services, which give them independence in life, but also to form a community within which they communicate with their peers. In case of any problems with the "consuming" function, senior citizens can ask the intermediary for help, what shall be depicted further in this article.

Service providers constitute another party to transactions realised with the usage of the IT product, and, as the term itself indicates, they provide services to senior citizens, thus performing the "offering" function in this model. In this function they specify their service offers, manage their own time or the time of their employees and/or volunteers, provide end users with the ability to book services, manage price lists and discounts, customise offers and confirm that the service will be performed.

IT product providers, in turn, have been assigned two basic functions in the presented model:

- Delivery,
- Maintenance.

The said functions are performed through developing the IT product in a particular programming language and through ensuring its broadly understood maintenance, covering not only the correction of errors but also product development and technological support for the users. The activities related to the 1st line technological support are conducted via the intermediary in this model. The aforementioned solution is applied in order to overcome the selected for PEST analysis barriers resulting from senior citizens' low digital competences and their problems with the perception of IT products. Furthermore, it should be noted that the "delivery" function is performed by the provider through delivering the IT product in the form of a service in the SaaS model. In practice, this means that the IT product, developed in compliance with the SOA principles is available via a browser, without the necessity to install client software. This responds to senior citizens' low digital competences, which were diagnosed with the application of

PEST analysis, and which manifest themselves in their lack of skills at computer administration, among other things.

The intermediary, in turn, has been assigned the following functions:

- Support,
- Mediation,
- Verification,
- Stimulation.

The intermediary is a stakeholder whose task is to respond to such factors selected in PEST analysis as: age-related mental, biological and social barriers in the perception of the IT product, senior citizens' low digital competences, distrust in new IT solutions and in service providers. The "support" function is performed by the intermediary through two basic activities:

- Providing a help-desk which offers ongoing technological and substantive assistance to senior citizens,
- Ensuring trainings for senior citizens, volunteers, and service providers.

The "mediation" function covers the relation between:

- Senior citizens, service providers and the IT provider, not only in the field of solving current problems but also of defining the needs which should be reflected in the IT product (functional analysis),
- Senior citizens and service providers in the area of financial transactions (negotiating the price, settlements of transactions) and of ordering the services on behalf of senior citizens (helpline)

The function designed to verify the quality offered by service providers covers both the preliminary assessment of their reliability before they are admitted to offer their services in the platform and the confirmation of the correctness of the services performed for senior citizens. With the usage of these tools, the intermediary builds broadly understood confidence in ordering services in this way.

The "stimulation" function, in turn, can be analysed in the context of cooperation with both senior citizens and service providers. The intermediary not only maintains contact with organizations whose task is to activate senior citizens and to bring them together, which is his way of attracting new end users, but he also actively searches for and attracts service providers which intend to reach senior citizens with their offer with the usage of the IT product. The intermediary can also attract individual volunteers who are interested in providing elderly people with their assistance. Furthermore, the intermediary stimulates senior citizens to communicate with one another within the IT product and to exchange opinions about service providers, thus enabling the users to form an active community.

The introduction of the intermediary, whose task is to organise broadly understood assistance for senior citizens, to actively stimulate them to use the IT product, to satisfy their everyday needs, to verify the quality of services and to supervise financial transactions, allows to promulgate the concept of the IT product among senior citizens and, first of all, to build trust in it. This is the way in which the barriers resulting from senior citizens' low digital competences and their distrust in technological and process innovations are overcome. What is also eliminated owing to the existence of the intermediary is the problem of dishonest service providers, which are monitored by the intermediary and end users (senior citizens) on an ongoing basis. While performing these organic functions, the intermediary can apply different communication channels offered by the IT product.

Basic functions performed by the stakeholders from macroenvironment are the following:

- To provide financial means applicable to create new IT solutions dedicated to senior citizens,
- To undertake measures to activate senior citizens to use such technological solutions,
- To amend legal regulations in an ongoing manner in order to adjust them to the changing social needs.

Current measures undertaken by the local, public and EU authorities support financially technological innovations designed to provide senior citizens with the means which will enable them to live independent lives. The examples of such initiatives have already been discussed in the Introduction. Owing to such programmes, the organisers of such social initiatives can obtain financing, and, consequently, they can offer their IT products for free. This is particularly important in Poland. The research conducted earlier clearly indicates that Polish senior citizens are reluctant to pay for the usage of applications due to their low incomes.

What also constitutes a significant source of financing of activities related to providing senior citizens with technological solutions is private sponsoring. Companies (State-owned companies or private companies) donate funds for such purposes according to the assumptions on the corporate social responsibility (CSR). The fact that companies participate in the life of the local community and that they engage themselves in social investments constitutes the root of their firm position in this community. Not only do they gain the hearts of the local population but also the trust of the local authorities.

It is also possible (at the early stage of the project) to obtain financing from venture capitals, which will receive their shares in such a project.

Private individuals also support such initiatives through small one-time donations (crowdfunding) made via such crowdfunding platforms as Kickstarter or PolakPotrafi.pl. By doing so, they express their support for such projects and they even receive some petty benefits (e.g. shares in the project, discounts for services, gadgets, etc.)

Activating elderly people in the context analysed in this article takes place through informing the target group about

the possibilities to take advantage of such applications. This might be done through organising different social events, workshops, lectures, trips etc. What should be presented during such meetings is the system itself together with its characteristics, possibilities, and basic functionalities. It is also important to promote an active lifestyle suitable for senior citizens and to show them paths of their prospective development (third age universities). Due to such measures, they will be more willing to benefit from different kinds of services ant it will be easier to persuade them to use the system described in this article.

This is the role performed mostly by NGOs and the media. Such actions should be also taken by care institutions, hospitals, church organisations and public figures. Due to their authority they will be able to recommend using such applications. What should also be promoted and rewarded are different forms of volunteering aimed at raising the awareness of senior citizens in the area of ordering services via the discussed system.

The Authors indicate legislative works, especially drafting legal provisions (both national and European regulations) as the third function performed by the stakeholders from macroenvironment of the IT product. The role of the said regulations is:

- To guarantee the security of transactions realised via the Internet,
- To protect personal data,
- To enable providers to offer services for elderly people in volunteering,
- To allow elderly people to work without losing their retirement pensions,
- To enable elderly people to offer mutual services (via the so called "time banks") and to settle their accounts (payments received for the work done) with the usage of virtual currency or barter.

The roles listed above and their functions constitute the basic structure of the model, yet, the possibility of its further extension or modification in non-standard cases is not excluded. The local or public authorities, for example, might perform the function consisting in activating elderly people even though the said function has been originally assigned only to the media and NGOs. Similarly, not only the intermediary but also service providers can boost the movement on the IT product web page through organising various promotional campaigns on their own web pages or during personal meetings with their clients.

IV.CONCLUSIONS AND FURTHER RESEARCH

The article presents the IT product delivery model dedicated to activate and support senior citizens in their independent everyday lives. The model has been developed on the basis of the strategic PEST analysis. The roles defined in this model and the functions performed by them constitute the response to the factors selected in PEST analysis and are designed to indicate the most important elements of the efficient delivery of the IT product which is suitable for the needs of the analysed target group. The Authors have also specified the strength of impact which particular roles and functions have by locating them in the proper layers of the onion model.

On the basis of the conducted research the following conclusions have been formulated:

- The presented model is of a comprehensive character and it covers such elements as the product, roles performed by particular stakeholders from micro- and macroenvironment and their functions,
- The elements of the model have been defined in such a way that they constitute a response to the factors selected during PEST analysis and if the factor is negative, they constitute an antidote which is capable of eliminating its influence,
- Particular roles have different strength/form of impact - they can influence the IT product interacting with it directly (delivery, maintenance, consuming/offering) or indirectly through creating favourable conditions for the usage of such solutions,
- Among the indicated roles which interact directly with the product, the role of the intermediary has been emphasised as he mediates between senior citizens and the other roles in the model and he facilitates senior citizens' usage of the IT product (eliminating age-related mental, biological and social barriers),
- The application of the suggested IT product delivery model constitutes an opportunity to reach the target group in the most efficient manner and to eliminate their individual limitations as well as the limitations resulting from the environment as identified in PEST analysis.

The research presented in this article does not cover all of the issues related to delivering IT products designed to increase the senior citizens' quality of life. On the one hand, what seems to be of particular importance is to deepen the analysis of all the stakeholders engaged in the process of the IT product delivery to the market, but on the other, the Authors acknowledge the need to model the functional structure of the product itself. Future papers authored by the research team shall be devoted to the aforementioned problems.

REFERENCES

[1] GUS, Prognoza ludności na lata 2014-2050, *Studia i analizy statystyczne*, 2014, Retrieved April 9, 2016, from: https://www.mpips.gov.pl.

[2] Active and Assisted Living Programme, Retrieved April 9, 2016, from: http://www.aal-europe.eu

[3] The Consumers, Health, Agriculture and Food Executive Agency, Call 2016, Retrieved April 9, 2016, from

http://ec.europa.eu/chafea/documents/health/hp-pj-2016-call-text_en.pdf
[4] MPiPS, Rządowy Program na rzecz Aktywności Społecznej Osób Starszych na lata 2014–2020, Retrieved April 9, 2016, from: https://www.mpips.gov.pl.
[5] Mindtools, PEST Analysis. Identifying "Big Picture" Opportunities and Threats, Retrieved April 9, 2016, from: http://www.mindtools.com.
[6] R.B. Duncan, Characteristics of organizational environments and perceived environmental uncertainty. *Administrative Science Quarterly* 17(3), 1972.
[7] D. Ward, E. Rivani, *An Overview of Strategy Development Models and the Ward-Rivani Model,* Economics Working Papers, 2005.
[8] K. Siau, R. Chiang, B. C. Hardgrave, *Systems Analysis and Design: People, Processes, and Projects,* Routledge, 2015, p.61
[9] M. E. Porter, *Competitive Strategy: Techniques for Analyzing Industries and Competitors,* 1st Edition, The Free Press, New York 1998
[10] B. Butryn, W. Gryncewicz, R. Kutera, M. Leszczyńska "The Application of PEST Analysis to the Creation of the Profile of an IT Product Designed to Activate and Support Senior Citizens in Poland", in *Proceedings of 9th International Symposium on e-Health Services and Technologies (EHST 2015),* Rhodes; 09/2015, pp. 109-115.
[11] D. Jelonek, C. Stępniak, T. Turek and L. Ziora, "Identification of mental barriers in the implementation of cloud computing in the SMEs in Poland," *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on,* Warsaw, 2014, pp. 1251-1258.

14^{th} Conference on Advanced Information Technologies for Management

WE are pleased to invite you to participate in the 14^{th} edition of Conference on "Advanced Information Technologies for Management AITM'16". 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

- 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
- **Atemezing, Ghislain,** Mondeca, Paris, France
- **Banaszak, Zbigniew,** Warsaw University of Technology, Poland
- **Bobkowska, Anna,** Gdansk University of Technology, Poland
- **Brown, Kenneth,** Communigram SA, France
- **Bruzda, Jaonna,** Nicolaus Copernicus University, Poland
- **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
- **Fosner, Maja,** Faculty of Logistics, University of Maribor, Slovenia
- **Franczyk, Bogdan,** University of Leipzig, Germany
- **Gontar, Beata,** Uniwersity of Lodz, Poland
- **Gontar, Zbigniew,** University of Lodz, Poland
- **Hartványi, Tamás,** Széchenyi István University, Hungary
- **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
- **Korczak, Jerzy,** Wrocław University of Economics, Poland
- **Kowalczyk, Ryszard,** Swinburne University of Technology, Melbourne, Victoria, Australia
- **Kozak, Karol,** TUD, Germany
- **Križanová, Anna,** University of Zilina, Slovakia
- **Langviniene, Neringa,** Kaunas University of Technology, Lithuania
- **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
- **Lim, Ming K.,** University of Derby, United Kingdom
- **Ludwig, André,** University of Leipzig, Germany
- **Magoni, Damien,** University of Bordeaux – LaBRI, France
- **Matulewski, Marek,** Poznań School of Logistics, Poland
- **Michalak, Krzysztof,** Wroclaw University of Economics, Poland
- **Montemanni, Roberto,** University of Applied Sciences of Southern Switzerland, Switzerland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland
- **Pamuła, Anna,** University of Łódź, Poland

- **Pankowska, Malgorzata,** University of Economics in Katowice, Poland
- **Patasiene, Irena,** Kaunas University of Technology, Lithuania
- **Pawełoszek, Ilona,** Częstochowa Univeristy of Technology
- **Quirin, Arnaud,** University of Vigo
- **Rakovska, Eva,** University of Economics in Bratislava, Slovakia
- **Ricci, Stefano,** Sapienza University of Rome, Italy
- **Rot, Artur,** Wroclaw University of Economics, Poland
- **Shinkevich, Aleksej Ivanovich,** Kazan National Research Technological University, Russia
- **Sitek, Pawel,** Kielce University of Technology, Poland
- **Speranza, Grazia,** University of Brescia, Italy
- **Stanek, Stanislaw,** 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,** University of Szczecin, Poland
- **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. 205–210
DOI: 10.15439/2016F465
ACSIS, Vol. 9. ISSN 2300-5963

Finding an Optimal Team

Michał Okulewicz
Warsaw University of Technology,
Faculty of Mathematics and Information Science,
Koszykowa 75, 00-662 Warsaw, Poland
M.Okulewicz@mini.pw.edu.pl

***Abstract*—This article proposes a metaheuristic optimization/social simulation approach to find the optimal team for a given type of the project. The quality of the team is assessed in a black-box optimization environment, where the optimized function acts as a metaphor of the project to be completed within the certain time limit (number of fitness function evaluations) and each fitness function evaluation is considered to be a metaphor of a unit task. The employees in a team are modeled according to the Belbin's Team Roles and the Particle Swarm Optimization (PSO) is used as a teamwork framework algorithm, while Evolutionary Algorithm (EA) as an algorithm for controlling the set of Team Roles for team members and leaders. This approach has been tested in a scenario of a simulated self-organizing team, where each employee decides about his own actions. The results from the performed simulation suggest, that such teams perform best if their leader is one of the actual work-oriented roles. Additionally, some projects required significantly different set of roles than the average team, resulting in improvement of the specialized team's performance over that of the average team.**

I. Introduction

IN 1981 and 1988 Raymond Meredith Belbin proposed his Team Roles theory [1]. This theory identifies nine clusters of people's behaviors expressed during the cooperation with other people. The clusters are build upon well-known psychometric factors and an individual's temperament assessment. Although the initial way of assigning team roles for the people has been criticized [2] the theory has been commercially successful, resulting in emergence of the Belbin Associates (http://www.belbin.com/).

From the author's personal team management experience, this theory has proven quite useful for choosing proper tasks assignment and making accurate tasks justification, especially in the lack of other motivation methods.

An automatic method of finding a group of people with such set of Team Roles, whom could efficiently complete a given type of project (easy/repeatable, with some obstacles but a known general method of approach, high-risk research and development etc.), would greatly lower the risk and cost of doing such project. Such method could be also combined with a methods of automatic project planning [3], [4] or reusing information from previous projects [5], providing a set of decision support systems for project managers. Although such an approach will always be prone to the errors resulting from the simplicity of behavior modeling, it is still useful because of the possibility of simulating a vast number of hypothetical situations [6] and observing the recurring patterns in the proposed team structure. Such a simple model allows for capturing the general features of the team member behaviors with connection to the problem (e.g. analyzing the alternatives, following the state-of-the-art approach). Other applications of virtual agents simulation results used for improving business processes can be found in [7].

As the Particle Swarm Optimization (PSO) has been originally proposed as a social simulation framework, this article proposes a generalization of the PSO particle, which would act as a metaphor of an employee.

The rest of the paper is organized as follows. Section II describes the generalized PSO particle and the generalized PSO as the team simulation framework. Section III presents Evolutionary Algorithm (EA) as a method of changing team members and leaders Team Roles. Subsequently, section IV defines the Team Roles from the Belbin's Theory. Sections V, VI and VII describe experiments setup and their results. Finally, section VIII concludes the paper.

II. Particle Swarm Optimization: Teamwork Framework

As already mentioned, Particle Swarm Optimization (PSO) has been initially designed as an algorithm mimicking simple social behavior [8], quite similar to Reynold's boids [9]. Also some recent research presents connection between complex network analysis and particle swarm behavior [10]. Therefore, it is well-suited to be interpreted as a sort of a multi-agent system providing a good framework for such simulated environment. Please note, that in the rest of the article the generalized PSO *particles* will be referred to as *employees*, if it will be more meaningful in the given context.

In order to allow *employees* to incorporate a different set of actions, modeling the behavior of an *employee* acting in a certain team role, the velocity update formula of the Standard PSO [11], has been generalized from choosing the best particle among particle's neighbors to a various aggregation formulas (see Table I). This could result in a behaviors of particles like in Charged Particle Swarm Optimization [12] (in order to incorporate possibility of repulse [13] from certain individuals) and Civilization Algorithm / PSO with charisma (in order to incorporate possibility of special treatment of the information from the team leader):

$$\begin{aligned} v \quad = \quad & \omega v + \\ & c_1 r_1 (x_{Best} - x) + \\ & (min_{c_2} + c_2) r_2 (x_{NeighborAggregation} - x) \end{aligned} \tag{1}$$

$$P(X \in Neighbors(Y)) = \begin{cases} 1, & X = Y \\ \dfrac{p_{ts}(X) + p_{tl}(Y)}{2}, & Team(X) = Team(Y) \\ \dfrac{p_s(X) + p_l(Y)}{2}, & Team(X) \neq Team(Y) \\ \min(1, p_{ts}(X) + p_{tl}(Y)), & Team(X) = Team(Y) \wedge (Leader(X) \vee Leader(Y)) \\ \min(1, p_s(X) + p_l(Y)), & Team(X) \neq Team(Y) \wedge Leader(X) \wedge Leader(Y) \end{cases}$$

Fig. 1. The probabilities that X would be a neighbor of Y in a given iteration.

Where:

- v is a velocity vector of the particle,
- x is a current location of the particle,
- x_{Best} is a best location visited by the particle,
- $x_{NeighborAggregation}$ is a location computed (aggregated) on the base of information from all the neighbors of the particle (typically just the best location),

Additionally:

- ω is an inertia coefficient,
- c_1 is a local attraction factor,
- c_2 (with additional min_{c_2} bias) is a neighbor attraction (could be used as repulsion) factor.
- r_1, r_2 are vectors of uniformly distributed random variables.

The communication topology is a random one, with four probabilities defined for each role and a team leader modifier. The possibility of communication between two *employees* is established independently in each iteration, following the four probabilities:

- p_{ts} a probability of being neighbor of a teammate (speaking),
- p_{tl} a probability of neighboring a teammate (listening),
- p_s a probability of being neighbor of an employee from another team (speaking),
- p_l a probability of neighboring an employee from another team (listening).

In order to reduce the number of parameters of the simulation, 5 levels of probabilities have been introduced:

1) $p_{HIGH} = 0.8$
2) $p_{TYPICAL} = 0.5$
3) $p_{LOW} = 0.1$
4) $p_{VERY\_LOW} = 0.01$
5) $p_{ZERO} = 0.0$

In addition, if a given employee is a leader he will have higher probabilities of communicating with other leaders and his team members. The detailed rules for computing probability of successful communication between two *employees* are presented in Fig. 1.

III. Evolutionary Algorithm: Team Management Framework

For managing the teams an Evolutionary Algorithm (EA) has been used (thus creating a meta-optimization like algorithm, although used for creating a certain Team Roles set rather than just optimizing a function).

The **fitness function** for the EA is a median of the difference between the achieved value of the function and the target value of the function within the limited budget of function evaluations on 15 different instances of the same function. The fitness function for the team is computed from the best values achieved by the members of this team.

The **cross-over operator** of the EA switches the Team Role of the team leader from the first parent with the Team Role of the team leader from the randomly chosen second parent, thus creating the child team.

The **mutation operator** of the EA randomly changes a team member's Role into another Role.

The **selection operator** is based on the tournament selection between the parents and the offspring in order to maintain high diversity of the population. Please note, that the offspring could be result of the cross-over operation or mutation. In the case of cross-over randomly chosen child would compete with first of the parents. In the case of the mutation the cloned mutant will compete with the original individual.

IV. Team Roles Models

In this section, a short description of each of the nine Team Roles will be given, summarized with the set of parameters (c_1, c_2, neighbor aggregation method, communication probabilities) used for modeling behavior for each role (see Table I). Each role description starts with a short quote from the http://www.belbin.com website. For a detailed descriptions please refer to the works of the R.M.Belbin [14]. After that short description each agent would be described in terms of communication abilities and methods of processing the received information.

A. Plant

> *The first Team Role to be identified was the Plant (PL). The role was so-called because one such individual was planted in each team. They tended to be highly creative and good at solving problems in unconventional ways.* [15]

PL is an individual worker with unorthodox ideas, but possibly some communication issues. PL works in isolation and tries to find new (uncommon) ways to solve the problems within the project. As a particle it acts as a charged particle in CPSO algorithm, repulsing from the average location of the best known results, thus searching for completely new solutions.

TABLE I
SUMMARY OF TEAM ROLES PARAMETERS

Team role	c_1	min_{c_2}	c_2	Aggregation method	p_{ts}	p_{tl}	p_s	p_l
Plant	1.4	-1.4	1.4	Average location	LOW	TYPICAL	ZERO	ZERO
Monitor Evaluator	0.0	0.9	0.2	Promising and not explored cluster centers	TYPICAL	TYPICAL	ZERO	ZERO
Coordinator	0.0	0.9	0.2	Subsequent cluster centers	TYPICAL	TYPICAL	LOW	LOW
Resource Investigator	0.0	0.9	0.2	Max distance	HIGH	HIGH	TYPICAL	TYPICAL
Implementer	1.4	0.0	1.4	Best neighbour	TYPICAL	TYPICAL	ZERO	ZERO
Completer Finisher	0.0	0.9	0.2	Best neighbour	LOW	VERY_LOW	ZERO	ZERO
Teamworker	0.0	0.0	1.4	Average location	TYPICAL	TYPICAL	LOW	LOW
Shaper	1.4	0.0	1.4	Best neighbour	HIGH	HIGH	LOW	LOW
Specialist	0.0	0.9	0.2	Function approximation	TYPICAL	TYPICAL	ZERO	ZERO

B. Monitor Evaluator

> *The Monitor Evaluator (ME) was needed to provide a logical eye, make impartial judgements where required and to weigh up the team's options in a dispassionate way.* [15]

ME greatest ability is to separate facts from opinions and assess situation without emotional biases. MEs focus on elaborating on all the plausible alternatives leading to achieving project's objectives.

As a particle, ME visits the locations which seem to be not explored enough, while having quite a good overall fitness function value. ME uses the UCB1 [16] approach in continuous problem by clustering the samples gathered by other particles, and computing the average value of the fitness function combined with the size (measured in number of samples) of the cluster. It uses such evaluation to choose which area should be explored, thus maintaining an exploitation-exploration balance. This behavior could also be looked upon as a sharing mechanism known from evolutionary approach, where fitness function quality is divided by the number of nearby specimen.

C. Coordinator

> *Co-ordinators (CO) were needed to focus on the team's objectives, draw out team members and delegate work appropriately.* [15]

CO's abilities concentrate around proper work division and tasks delegation.

As a particle CO explores the area of each of the samples' clusters in a subsequent manner. CO behavior is closely related to the work of the ME, but CO focuses its attention on each of the clusters regardless of its average function value.

D. Resource Investigator

> *When the team was at risk of becoming isolated and inwardly-focused, Resource Investigators (RI) provided inside knowledge on the opposition and made sure that the team's idea would carry to the world outside the team.* [15]

RI gathers information about other teams results (through high probability of communicating with the members of other teams).

As a particle, RI explores the promising areas which are the furthest from its current location.

E. Implementer

> *Implementers (IMP) were needed to plan a practical, workable strategy and carry it out as efficiently as possible.* [15]

IMPs might be looked upon as the backbone member of a team. In the simulation, IMP acts as a standard PSO particle, which simulates following a most natural strategy balancing a choice between best external information and best personal experience.

F. Completer Finisher

> *Completer Finishers (CF) were most effectively used at the end of a task, to "polish" and scrutinise the work for errors, subjecting it to the highest standards of quality control.* [15]

CF likes to work on a task until it is properly finished. Therefore CF's listening ability is low, as CF is concentrated on his work (on the other hand, CF provides information about the progress). As a particle, CF is attracted to the best neighbor location, but due to low probability of being informed about a new location it explores one area for a longer period of time.

Due to ω factor it is expected to oscillate around a promising location (until finding a better one) with the smaller steps at each iteration, therefore acting similar to a Variable Neighborhood Search algorithm.

G. Teamworker

> *Teamworkers (TW) helped the team to gel, using their versatility to identify the work required and complete it on behalf of the team.* [15]

TW tries to help people with their work. Although TW can communicate easily, such employee will not necessarily engage in a work related conversation.

As a particle TW is attracted by all of its neighbours working like a particle in Fully Informed PSO algorithm [17].

H. Shaper

> *Challenging individuals, known as Shapers (SH), provided the necessary drive to ensure that the team kept moving and did not lose focus or momentum.* [15]

SH tries to finish project as fast as possible (possibly even at the cost of its quality). SHs like to influence the way other

TABLE II
AVERAGE FREQUENCY OF THE ROLES WHILE ACTING AS A TEAM MEMBER AND TEAM LEADER

Function	CF	CO	IMP	ME	PL	RI	SH	SP	TW
Team leader	0.45	0.04	0.11	0.03	0.03	0.05	0.12	0.08	0.08
Team member	0.23	0.09	0.12	0.09	0.08	0.09	0.10	0.11	0.09

TABLE III
AVERAGE FREQUENCY OF THE ROLES WHILE ACTING AS A TEAM LEADER FOR DIFFERENT FITNESS FUNCTIONS. THE FUNCTIONS ARE DIVIDED INTO 5 GROUPS: SEPARABLE, WITH LOW OR MODERATE CONDITIONING, WITH HIGH CONDITIONING, MULTI-MODAL WITH GLOBAL STRUCTURE, MULTI-MODAL WITH WEAK GLOBAL STRUCTURE [18]

Optimized function	CF	CO	IMP	ME	PL	RI	SH	SP	TW
f1	0.24	0.12	0.10	0.10	0.00	0.06	0.06	**0.30**	0.02
f2	**0.38**	0.00	0.20	0.00	0.00	0.02	0.30	0.10	0.00
f3	**0.30**	0.02	0.08	0.00	0.04	0.08	0.28	0.10	0.10
f4	**0.42**	0.02	0.14	0.00	0.08	0.02	0.30	0.02	0.00
f5	0.12	0.00	0.16	0.00	0.00	0.10	**0.52**	0.08	0.02
f6	**0.72**	0.00	0.10	0.00	0.00	0.00	0.18	0.00	0.00
f7	**0.48**	0.00	0.14	0.00	0.00	0.00	0.18	0.10	0.10
f8	**0.76**	0.02	0.04	0.12	0.00	0.00	0.00	0.04	0.02
f9	**0.54**	0.00	0.02	0.08	0.08	0.14	0.06	0.08	0.00
f10	**0.60**	0.16	0.06	0.00	0.00	0.00	0.00	0.18	0.00
f11	**0.46**	0.06	0.00	0.00	0.02	0.04	0.00	0.16	0.26
f12	**0.46**	0.12	0.22	0.02	0.00	0.00	0.16	0.00	0.02
f13	**0.50**	0.10	0.12	0.00	0.10	0.04	0.00	0.00	0.14
f14	**0.52**	0.00	0.02	0.02	0.12	0.10	0.16	0.00	0.06
f15	**0.54**	0.00	0.14	0.00	0.00	0.02	0.08	0.10	0.12
f16	**0.70**	0.02	0.18	0.00	0.00	0.00	0.00	0.10	0.00
f17	**0.50**	0.00	0.00	0.08	0.10	0.02	0.04	0.22	0.04
f18	**0.56**	0.08	0.08	0.02	0.02	0.14	0.00	0.00	0.10
f19	0.06	0.00	0.22	0.10	0.04	0.06	0.00	0.02	**0.50**
f20	**0.42**	0.00	0.22	0.00	0.10	0.00	0.20	0.00	0.06
f21	**0.46**	0.06	0.14	0.06	0.06	0.04	0.00	0.12	0.06
f22	**0.26**	0.18	0.18	0.00	0.00	0.04	0.18	0.14	0.02
f23	**0.40**	0.00	0.04	0.08	0.00	0.26	0.00	0.04	0.18
f24	**0.51**	0.00	0.16	0.04	0.00	0.00	0.16	0.00	0.13

people work (if SH thinks, that the project would benefit from this).

As a particle SH acts exactly as IMP, but will have higher communication probability, which will speed up the convergence of the particles, thus resulting in finishing work earlier (although possibly in a much worse than optimal location).

I. Specialist

> *It was only after the initial research had been completed that the ninth Team Role, Specialist (SP) emerged. In the real world, the value of an individual with in-depth knowledge of a key area came to be recognized as yet another essential team contribution.* [15]

SP takes pride in being an expert and does not perform very well in cooperation, therefore as a particle SP observes the gathered samples and builds a set of linear models approximating promising areas of the optimized functions and then explores the area near the peak of approximating square functions.

V. ASSESSMENT ENVIRONMENT

As an assessment environment the GECCO (since 2009) and CEC (since 2015) BlackBox Optimization Benchmark set [18] has been used. It consists of 24 functions divided into 5 different categories, which would allow to have comparison for a different project difficulty levels.

This way, additional objective information on the quality of the team's performance has been obtained by comparing it with standard optimization algorithms (although the research does not focus on this part).

VI. TESTS

The training of the teams has been done with 10 repetitions of the experiment with 9 teams with 10 teamembers each with 1000 PSO iterations and 20 EA iterations on the set of all 24 benchmark functions (15 instances of each function).

The results of the training provided a frequency of each role acting as a team leader or a team member for each of the functions in the final population within the EA algorithm. The specialized team's were tested against the team generated from the average frequency of roles from all the functions.

VII. RESULTS

The results of the average roles frequency found within the training phase are presented in Table II. It can be seen that the **Completer Finisher** (CO) has been chosen as the most frequent Team Role for both the team member and the team

TABLE IV
AVERAGE FREQUENCY OF THE ROLES WHILE ACTING AS A TEAM MEMBER FOR DIFFERENT FITNESS FUNCTIONS. THE FUNCTIONS ARE DIVIDED INTO 5 GROUPS: SEPARABLE, WITH LOW OR MODERATE CONDITIONING, WITH HIGH CONDITIONING, MULTI-MODAL WITH GLOBAL STRUCTURE, MULTI-MODAL WITH WEAK GLOBAL STRUCTURE [18]

Optimized function	CF	CO	IMP	ME	PL	RI	SH	SP	TW
f1	0.13	0.08	0.11	0.09	0.07	0.09	0.09	**0.27**	0.07
f2	**0.25**	0.06	0.16	0.08	0.07	0.06	0.18	0.08	0.07
f3	**0.26**	0.07	0.11	0.10	0.08	0.10	0.12	0.09	0.07
f4	**0.22**	0.09	0.14	0.08	0.08	0.09	0.15	0.08	0.07
f5	0.15	0.07	0.13	0.08	0.10	0.09	**0.19**	0.13	0.07
f6	**0.29**	0.07	0.14	0.08	0.10	0.06	0.14	0.07	0.05
f7	**0.26**	0.09	0.10	0.08	0.07	0.08	0.12	0.10	0.10
f8	**0.32**	0.06	0.11	0.12	0.08	0.07	0.08	0.09	0.08
f9	**0.29**	0.09	0.11	0.08	0.09	0.07	0.10	0.09	0.08
f10	**0.31**	0.10	0.09	0.09	0.08	0.07	0.06	0.12	0.09
f11	**0.27**	0.10	0.10	0.08	0.09	0.09	0.08	0.11	0.09
f12	**0.22**	0.10	0.15	0.07	0.07	0.09	0.10	0.10	0.08
f13	**0.26**	0.09	0.15	0.08	0.08	0.10	0.05	0.10	0.09
f14	**0.27**	0.07	0.09	0.09	0.10	0.10	0.10	0.08	0.09
f15	**0.26**	0.09	0.13	0.08	0.09	0.10	0.05	0.11	0.08
f16	**0.23**	0.09	0.12	0.08	0.07	0.11	0.10	0.10	0.10
f17	**0.21**	0.09	0.12	0.12	0.12	0.09	0.08	0.10	0.07
f18	**0.24**	0.10	0.12	0.07	0.08	0.11	0.08	0.10	0.10
f19	0.18	0.09	0.13	0.08	0.06	0.06	0.10	0.12	**0.19**
f20	**0.19**	0.08	0.11	0.08	0.11	0.09	0.14	0.09	0.10
f21	**0.24**	0.09	0.11	0.06	0.08	0.09	0.09	0.14	0.09
f22	**0.18**	0.13	0.13	0.09	0.07	0.10	0.11	0.11	0.09
f23	**0.19**	0.07	0.12	0.07	0.08	0.14	0.09	0.08	0.16
f24	**0.22**	0.09	0.13	0.08	0.05	0.09	0.11	0.08	0.14

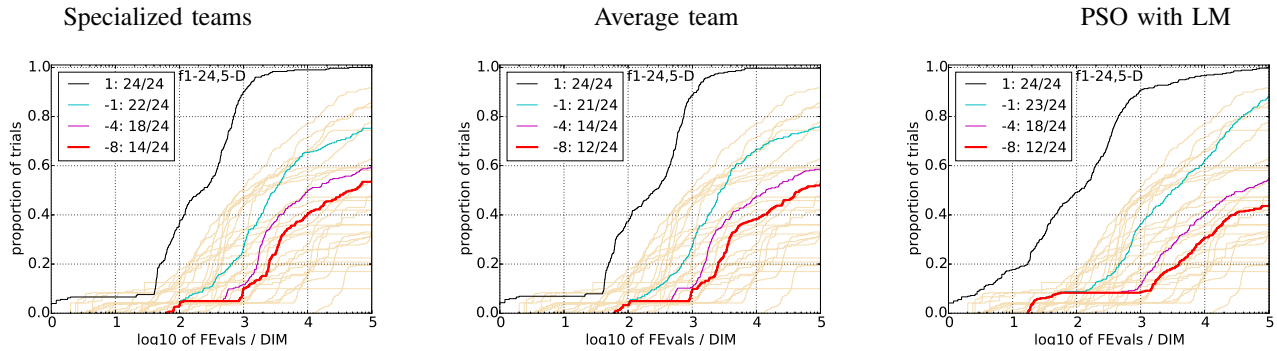


Fig. 2. Figures presenting the performance of the teams on a 5 dimension set of benchmark functions in terms of the fraction of runs (vertical axis) reaching certain optimization targets within the logarithm of certain number of function evaluations (horizontal axis). Additionally, a performance of a PSO algorithm enhanced with a linear model (approximating the optimized function with parabolas) is given.

leader. In the case of the team leaders the **Shapers** (SH) and **Implementers** (IMP) were chosen more frequently then the other 6 Team Roles, while in the case of the team members there seems to be no other significant distinction between average occurrences in the Team Roles except for the already mentioned CF. As can be seen from Tables III and IV the roles in the teams differ between the functions and the leader is always chosen among the most frequent of them.

The performance of the teams constructed from the average frequency of roles vs. specialized frequency of roles vs. PSO algorithm enhanced with a linear model with square function approximation is presented on Fig. 2. Although the performance in terms of the fraction of tries reaching 10^{1}, 10^{-1}, 10^{-4} and 10^{-8} targets is practically indistinguishable, on the other hand the number of types of functions for which certain optimization targets have been reached is higher for the specialized teams.

VIII. CONCLUSIONS

The proposed simulation showed the bias for team leaders towards the focused on the goal and hard-working team roles: Completer Finisher, Shaper, Implementer, Specialist and Teamworker. The results suggest that such leaders might lead to a best performance of a self-organizing team (as in the test framework each *employee* decided on its own about the task SP is going to perform). Additionally, from the average team members frequencies, the necessity for the representative of each of the roles among the team members has been stated. Some of the results, definitely still need a closer examination as they might be a result of a too simple implementation

(especially of the Monitor Evaluator and Coordinator roles). Also lack of discrimination between reported and real best position might lead to small or improper impact on the teamwork by the Monitor Evaluator, Coordinator and Resource Investigator roles. It proved beneficial not only to find a proper balance of roles and most likely roles to act as a team leaders, but also to find teams with that balance specialized for the given problems, as the specialized teams performed slightly better then the average teams (see the number of functions for which the optimization goals were achieved on the Fig. 2).

In addition it is important to observe, that the proposed algorithm achieved quite good results as an optimization algorithm. It has been able to find the optimum value with difference from the target optimum lower then 10^{-1} in case of almost all the functions and lower then 10^{-8} in case of more than the half of them, achieving better results for the hardest 10^{-8} goal then a reference enhanced PSO. It should also be noted, that the algorithm has been tuned on the basis of performance with around 4.25^{10} fitness function evaluations budget, which can be observed in comparison to PSO performance (see Fig. 2).

Future work

The future work should consist of building a more advanced simulation environment. It would be beneficial for the actions and interactions of the artificial agents actions to be related to the actions needed to complete some abstract project which can be later executed by human agents team, constructed according to the simulations results.

Additionally, the *employees* models might take into the account such possibilities as:

- migration of employees within teams set (instead of just changing into another role),
- conflicts between the team roles approach to work (resulting in loosing team members and the information gathered by them),
- reporting different achieved results in order to direct other *employees* to search given locations,
- *employees* being a probabilities vector of acting as a given role rather than just one team role.

The studies on the test environment itself might include:

- using asynchronous particles with time limit instead of a synchronized swarm with iterations budget bound,
- managing the team roles with another type of fitness function (taking into account cooperation of teams).

REFERENCES

[1] R. M. Belbin, *Management teams: why they succeed or fail*, 1st ed. Routledge, 1981.

[2] A. Furnham, H. Steele, and D. Pendleton, "A psychometric assessment of the Belbin Team-Role Self-Perception Inventory," *Journal of Occupational and Organizational Psychology*, vol. 66, no. 3, pp. 245–257, 1993. doi: 10.1111/j.2044-8325.1993.tb00535.x. [Online]. Available: http://dx.doi.org/10.1111/j.2044-8325.1993.tb00535.x

[3] K. Walędzik, J. Mańdziuk, and S. Zadrożny, "Proactive and reactive risk-aware project scheduling," in *Computational Intelligence for Human-like Intelligence (CIHLI), 2014 IEEE Symposium on*. IEEE, 2014. doi: 10.1109/CIHLI.2014.7013392 pp. 94–101. [Online]. Available: http://dx.doi.org/10.1109/CIHLI.2014.7013392

[4] ——, "Risk-aware project scheduling for projects with varied risk levels," in *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE, 2015. doi: 10.1109/SSCI.2015.231 pp. 1642–1649. [Online]. Available: http://dx.doi.org/10.1109/SSCI.2015.231

[5] Ł. Osuszek and S. Stanek, "Case based reasoning as an improvement of decision making and case processing in adaptive case management systems." in *Position Papers of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 6. PTI, 2015. doi: 10.15439/2015F61 pp. 217–223. [Online]. Available: http://dx.doi.org/10.15439/2015F61

[6] K. M. Carley, "Computational organizational science and organizational engineering," *Simulation Modelling Practice and Theory*, vol. 10, no. 57, pp. 253 – 269, 2002. doi: http://dx.doi.org/10.1016/S1569-190X(02)00119-3 Organisational Processes. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1569190X02001193

[7] M. Żytniewski, A. Sołtysik, A. Sołtysik-Piorunkiewicz, and B. Kopka, "Modeling of software agents' societies in knowledge-based organizations. the results of the study." in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015. doi: 10.15439/2015F216 pp. 1603–1610. [Online]. Available: http://dx.doi.org/10.15439/2015F216

[8] R. C. Eberhart, J. Kennedy *et al.*, "A new optimizer using particle swarm theory," in *Proceedings of the sixth international symposium on micro machine and human science*, vol. 1. New York, NY, 1995. doi: 10.1109/MHS.1995.494215 pp. 39–43. [Online]. Available: http://dx.doi.org/10.1109/MHS.1995.494215

[9] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *ACM SIGGRAPH computer graphics*, vol. 21, no. 4. ACM, 1987. doi: 10.1145/37401.37406 pp. 25–34. [Online]. Available: http://dx.doi.org/10.1145/37401.37406

[10] R. Šenkerık, M. Pluhácek, A. Viktorin, and J. Janoštık, "On the application of complex network analysis for metaheuristics," in *7th BIOMA Conference*, 2016, pp. 201–213. [Online]. Available: http://bioma.ijs.si/proceedings/2016/14%20-%20On%20the%20Application%20of%20Complex%20Network%20Analysis%20for%20Metaheuristics.pdf

[11] M. Clerc, "Standard Particle Swarm Optimization. From 2006 to 2011," 09 2012. [Online]. Available: http://clerc.maurice.free.fr/pso/SPSO_descriptions.pdf

[12] T. M. Blackwell and P. J. Bentley, "Dynamic search with charged swarms," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. ISBN 1-55860-878-8 pp. 19–26. [Online]. Available: http://dl.acm.org/citation.cfm?id=646205.682961

[13] B. L. J. Zhang, C. Fan and F. Shi, "An Improved Particle Swarm Optimization Based on Repulsion Factor," *Open Journal of Applied Sciences*, vol. 2, no. 4B, pp. 112–115, 2012. doi: 10.4236/ojapps.2012.24B027. [Online]. Available: http://dx.doi.org/10.4236/ojapps.2012.24B027

[14] R. M. Belbin, *Team roles at work*, 2nd ed. Routledge, 2012.

[15] BELBIN Associates, "Belbin Team Roles," 05 2015. [Online]. Available: http://www.belbin.com/

[16] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002. doi: 10.1023/A:1013689704352. [Online]. Available: http://dx.doi.org/10.1023/A:1013689704352

[17] R. Mendes, J. Kennedy, and J. Neves, "The fully informed particle swarm: simpler, maybe better," *Evolutionary Computation, IEEE Transactions on*, vol. 8, no. 3, pp. 204–210, 2004. doi: 10.1109/TEVC.2004.826074. [Online]. Available: http://dx.doi.org/10.1109/TEVC.2004.826074

[18] A. Auger, N. Hansen, M. Schoenauer, O. A. Elhara, A. Atamna, D. Brockhoff, A. Liefooghe, T.-D. Tran, R. Datta, R. L. Riche, E. Touboul, X. Bay, X. Delorme, D. Fongang-Fongang, H. Mohammadi, D. Villanueva, G. Rudolph, M. Preuss, H. Trautmann, O. Mersmann, B. Bischl, and T. Wagner, "COmparing Continuous Optimisers: COCO," 01 2015. [Online]. Available: http://coco.gforge.inria.fr/

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 211–218
DOI: 10.15439/2016F585
ACSIS, Vol. 9. ISSN 2300-5963

Megamodel-based Management of Dynamic Tool Integration in Complex Software Systems

El Hadji Bassirou TOURE, Ibrahima FALL, Alassane BAH, Mamadou Samba CAMARA
Institut de Recherche pour le Développement (IRD),
École Supérieure Polytechnique (ESP),
Université Cheikh Anta Diop de Dakar (UCAD), Sénégal
Email: {bassirou.toure, ibrahima.fall}@esp.sn
{alassane.bah, mamadou.camara}@ucad.edu.sn

***Abstract*—The development of complex software systems is more and more based on the composition and integration of autonomous component systems. This can be done either statically (proactive approach) at development-time or dynamically through a reactive approach in which a new composite system can possibly be created on-demand and/or at run-time from existing systems. With the aim of constructing and managing such complex and reactive software systems, we propose a megamodel-based environment supporting dynamic tool integration. Such an environment must therefore be consistent at any time (i.e. before, during and after an integration) and should also have to exhibit some self-* properties (such self management, self-healing and self-configuration). In order to meet these challenges we propose the use of Hoare's Axiomatic Semantics and some inference rules to maintain the integrity of the megamodel and its components. For that we have defined a formal-safe execution as well as an execution semantic for each operation likely to modify the megamodel contents.**

***Index Terms*—Dynamic Tool Integration, Complex System, MDE, Megamodeling, Axiomatic Semantics, Verification.**

I. INTRODUCTION

NOWADAYS software systems, such *Component-Based Software Engineering (CBSE)*, are generally composite systems which are becoming more and more complex. These systems are usually built on top of a generic platform by plugging into it many different software components and tools.

Such platforms are often extended or specialized for a given domain by respectively adding or removing one or several software components. This can be done either statically in a *proactive approach* at development-time, or dynamically at run-time through a *reactive approach*. We will present in the following some advantages and drawbacks for both proactive and reactive integration approach. A comparative study of the two approaches are extensively developed in [10] to which we refer the reader for further information.

Proactive approach

A proactive integration is an approach in which application designers implement a new application manually by designing correspondence or composition rules describing the interaction patterns between its components. Such an approach is set up when the architecture is still under development, namely at design or development time.

Limits

It complicates the problem of ensuring consistency in the software systems and is severely limited in flexibility;

Advantages

It supports more powerful integration methods and ensures that the adaptation will not produce anomalous behavior. This is due the fact that a *proactive approach* try to identify the potential effect of making a change even before the change has actually been made. It might also prohibit certain changes that would otherwise lead to unexpected behavior [11].

Reactive approach

In a dynamic (or reactive) integration a new composite system can possibly be created on-demand and/or at runtime from existing systems, considering the user preferences and the context to personalize the system maintenance process.

Limits

It is difficult to use traditional testing and formal verification techniques to check safety and other correctness properties.

Advantages

Dynamic integration allows software components (module behavior) and their interactions to be changed while modules are executing. Such a reactive approach starts to work at the moment the actual changes are being made, and typically try to resolve potential inconsistencies interactively. Moreover it may also enable a number of useful applications that could not be envisioned during development time. Therefore, such a kind of integration is suitable for end-user applications where available components are dynamic and users needs may be varying frequently.

Abstraction plays an important role in component-based programming, it is taken into account through encapsulation which ensures information hiding and independence between components [20]. To provide such an abstraction while building complex software systems, Model-Driven Engineering (MDE) seems to be a preferential solution. In fact, it is a "*recent*" Software Engineering (SE) discipline which promotes *models* as first class entities in the software system development and maintenance. In the MDE field, a *model* is defined as an artifact which consists of model elements, conforms to a specific *metamodel* and represents a given view of a system. A metamodel can be defined as a language that describes the various kinds of contained model elements and the way they are arranged, related and constrained [1]. However a system is often represented by various kinds of *interrelated models*.

Moreover, the architecture of a software system defines such a system in terms of components and interactions between them. MDE provides the concept of a megamodel as a building block for modeling in the large [3], thus has to hide fine-grained details that obscure understanding and focus on the "*big picture*", i.e. *system structure*, *interactions between models*, *assignment of models as parameters or results for model transformations*, *and so on*. Indeed, megamodeling offers the possibility to handle models (and metamodels) as first-class entities, to specify relationships between them and to navigate among them. Furthermore, by keeping track of all the heterogeneous modeling artifacts (models, metamodels, DSL, etc.) within a megamodel, all of them are treating as models. This results in a homogeneous infrastructure which enables the management of complex modeling artifacts [2].

Our purpose is to use *megamodels* for representing the components of an architecture and the interactions between them. Thereby to develop an integration management approach based on operations defined in the megamodel. As already stated a *megamodel* refers to a model that have models as its elements and that captures the interconnections between multiple models(*component models*) in the form of model operations, generally represented as model transformations (*global operation models*) [1]. All of these *models* will then be represented in a single runtime megamodel which will be handled as an execution environment or more simply as a (mega-)program which is updated with each new *(global) operation execution* [6]. The megamodel is therefore subjected to frequent dynamic changes which consist of either *adding* or *removing components*. However such changes must not violate the integrity of the megamodel and its constituent components. It is therefore necessary for one to be able to add or remove components while maintaining some kind of integrity of the entire system represented by a megamodel.

In order to meet these challenges we consider a megamodel as a program and use techniques for proving program correctness. These techniques are known as Hoare's axiomatic semantics and are used with some inference rules for checking the megamodel's consistency by defining, for each *global operation* likely to modify the megamodel contents, a formal and safe execution as well as an execution semantic which denotes the observable behavior of a program as it is executed.

The rest of the paper is organized as follows. Section II presents the problem statements. Section III is reserved for related works in which we present some papers that use megamodeling techniques. Section IV presents our approach of megamodel management of dynamic tool integration. Section V is reserved for the use of axiomatic semantics in order to provide a mean to check the megamodel's consistency. Section VI presents an example in which we illustrate the presented approach. Section VII concludes the paper and gives its future works.

II. Problem Statements

In the MDE vision, software development and management processes involve the creation and use of many related modeling artifacts which are becoming increasingly important. As a consequence, there is the need for efficient mechanisms to manage this constantly growing number of models which is due to many reasons, as those that follow [21]:

- Each viewpoint of the software is represented using a model with respect to the most adapted formalism (metamodel).
- Complex models need to be decomposed into smaller ones with different levels of abstraction.
- On behalf of the "*separation of concerns*", different models are created for different purposes.

Many aspects of model-management have been considered in the literature in which the concept of megamodels ([1], [5], [6], [15], etc.) and macromodels [21] have been proposed. However all of them, except in [15], focus on the management of *development models* whereas it could be very interesting to extend the use of models such as macromodels and megamodels for the management of *runtime artifacts* through the use of *runtime models*. Such runtime artifacts may include *component creation and destruction*, *exceptions/errors*, *operation inputs and output*, *components invocation operations*, *dynamic artifact types*, *dynamic component names*, and *so on*.

Otherwise, causal connection between models and represented systems means that each time one read the model, he gets the information representing the current system state, and similarly, each time one writes the model, the information he writes makes the proper system change [16]. Using models at runtime for specifying runtime artifacts have two main requirements. 1. the model as interrogated should provide up-to-date and exact information about the system. 2. if the model is causally connected, then adaptations can be made at the model level rather than at the system level.

The importance of the use of models at runtime has been extensively discussed in [20]. In that paper, several problems for which runtime models could be useful have been proposed without giving any piste about how to implement a solution. Among those problems, one can cite the support of semantic integration of software components represented through runtime models. For example, suppose that we have a user who expresses a request in order to merge two models into one. To achieve this, most of the model-based approaches such in [12], [13] and [14], use development models by putting the system offline, at first. Then they specify a set of correspondence or composition rules generally using the respective metamodels of the two input models. Finally they restart the system to apply changes in the system. Another solution which does not necessitate of putting down the system consists of describing directly the merging operation by reasoning on runtime models. In this case, when implementing such an approach, one has to recognize that the solution will never stay constant. That is, it could happen a situation where new models are introduced in the running system, and perhaps, some models to be removed from it. This continuous change necessitates the modeler to focus on how the system will react to those frequent changes and therefore how it will evolve

over time. An important challenge here is how to realize such a solution while maintaining some kind of integrity of the entire system. To achieve this, the approach should therefore implement a reactive environment to face changes in the running system. In order to satisfy users requests for example, such an environment should have to exhibit some self-* properties as self-management, self-healing and self-configuration.

The presented model-based approaches ([12], [13] and [14]) solve the problem of integration of software components by using the first solution through the use of development models, therefore such approaches can be considered as proactive integration. But we found no approach carrying out this problem using runtime models.

III. Related Works

In this section, some approaches of model management are presented. Such approaches use some specific models such as megamodels or macromodels to manage changes in a complex modeling environment. Megamodels and macromodels based-approaches provide a framework for an efficient creation, storage, access and execution of large amounts of modeling artifacts and their interconnections [5]. Indeed each of these approaches provides a mechanism to represent and control changes that occur in the megamodel through various techniques.

In [1], Bezivin and al. are experimenting through their *AMMP* (*Atlas Model Management Platform*) environment the need to consider separately the activities of modeling in the small and modeling in the large. A *megamodel* is considered as *a kind of registry that can be seen as a model which elements are models or refer to models*. Thus a megamodel will help the *AMMP* platform to know the available tools or services. Authors use megamodels (which provide a global view on models) for the support of model-driven software development by using it for model management. Megamodels are also applied to facilitate traceability between models and their elements.

In [4], authors present *MoScript* which is a megamodel agnostic platform and a textual DSL for accessing, querying and manipulating modeling artifacts represented in a megamodel. Several modeling tasks are performed using different kinds of operations which involve *operations without side effects* and *operations with side effects*. *Operations without side effects* (*QueryOp*, *TransformOp*, *ProjectOp*, *StateCheckOp*) are those that do not modify the megamodel contents. *Operations with side effects* (*SaveOp*, *RemoveOp*, *RegisterOp*) are operations that are likely to modify the megamodel contents. MoScript also allows to write queries that retrieve models from a repository or that register newly produced models back to the repository.

In [21], a *macromodel* is defined as *a model consisting of elements denoting models and links denoting intended relationships between these models with their internal details abstracted away*. For an efficient management of models, the author considers different kinds of modeling artifacts (*models and relationships*) which act on different layers (*orders of hierarchy*. Applications of the use of these different models could be twice : the consistency checking between constituent models and the inference of relations from other relations.

In [6], the author considers a *megamodel* as *a program in which the declaration and definition of models within a megamodel as statements of a model-based programming language*. Then the execution of a simple program composed of a sequence of such statements manipulates the contents of a megamodel. The important contribution of handling a megamodel as being a program is that it enables the prevention of typing errors during the execution of such programs. An typing error is for example the application of a function on arguments for which it was not defined.

In [15] authors propose the use of megamodels for the management of models at runtime. A runtime model provides a viewpoint, on a running software system, that is usually used for managing the system. Authors present a set of models, which have to be managed at runtime, as *Reflection models*, *Implementation models*, *Evaluation models*, *Change models*, *Monitoring models* and *Execution models*. In order to manage these models authors also propose a set of operations such *Update*, *Effects*, *Check for failures*, *Failure analysis rules*, *Repair strategies*. They propose the use of megamodels at runtime for both *navigation* and *automation*.

IV. Our approach : Use of MDE techniques for dynamic tool integration

Our motivation here is to use MDE techniques, particularly runtime models, in order to set up a megamodel-based environment supporting dynamic tool integration in a complex software system. In fact, our management of changes must ensure the consistency, correctness, and any desired properties of runtime change in the megamodel. In other words, our proposition for supporting dynamic integration should :

- Use a *megamodel* as a basis for model changes and management and which represents the dynamic architectural model of the global system;
- Provide inference rules for reasoning about and preventing changes from violating the *megamodel*'s integrity ;
- Help identify what models are likely to be targeted/modified, for example, by the insertion of a new model into the running system.

In this section we present at first the problem of tool integration, then we show how a tool is represented through various kinds of models and their interconnections in the megamodel. Finally we present our megamodel management and its constituent components.

A. Tool integration

A problem for tool integration is the general lack of standards for representing tools and their relationships. This has slowed the creation of fully integrated environments. In fact, one important barrier in current software engineering environments is the difficulty of integrating tools that address different aspects of the development process. As mentioned

in [8], integration is not a property of a single tool, but of its relationships with other elements in the environment. The key notion is the relationships between tools but also the properties of these relationships. These relationships can be represented through dimensions or types; a very consistent work on this purpose was attributed to Wassermann in [7] whose dimensions of tool integration are quoted by almost the totality of the other authors. The *platform dimension* for tool integration is concerned with the framework services that are commonly used by tools. *Presentation integration* deals with user interaction and *data integration* with data interchange. *Control integration* is concerned with interoperability between tools, and finally *process integration* refers to the different roles played by tools during a whole software process.

B. Tools as integrated models

In component-based programing a software system is conceived as a collection of several tools. Each tool contains some components which could encapsulate some services and provides or requires some operations from other tools. The components are always reached through the operations of the tool. They are said to be encapsulated in the tool. For each tool the problem is how to offer services, each one representing an integration dimension, to other tools and if there is the need how to use their services, in order to achieve a common goal. For that, it is necessary that each time a tool is added into the software system, it has to be registered as its supported services. Such a situation supposes to have a common view of what are a tool, its implemented services, operations it offers or requires etc. With respect to these requirements, we assume, as in [1], that *tools* will be handled as *models*.

Each tool consists of component models which could encapsulate some services. A service may correspond to any element of the Wasserman's integration dimension [7]. It can then be *data*, a *function*, a *process element*, *platform element* or *presentation element*. A tool may also provide or require some operations from other tools. Indeed, given two different tools, the interactions between two corresponding models come in the form of model operations (*global operation models*). A *global operation model* for example may also have input and output *parameters*, each *parameter* being itself considered as a *component model*. Thus a megamodel consists of *component models* and *global operation models*. A *component model* encapsulates artifacts which represent *services* of a tool. A *global operation model* can be seen as a type of an operation between component models. It therefore represents a model of future interactions (*global operation instances*) that connect some *component instances*. In other words, a *global operation model* defines one or more interaction rules, can be instantiated on *component instances* and allows to dynamically establish links between components. A *global operation model* can only be applied to component models already contained in the megamodel, and its results are *new component models* which are automatically added in the megamodel.

Manipulating a megamodel is then like programming where the megamodel acts as an environment, or more simply as a (mega-) program. It can be updated with each new *global operation execution*. However considering a megamodel as a program has already been proposed in [6]. In such contexts, the megamodel is subjected to frequent and dynamic changes which can due to the execution of operations (instructions) on components such the *introduction of new components*; the *recreation of failed components*; the *modifications of component interconnections*; the*change component operating parameters*, etc. A *global operation models* corresponds to a *set of operations* which are executed in response to changes related to the underlying system state changes. Indeed, after each execution of a global operation model, the megamodel has to be updated. Then, each component has the possibility/responsibility to update the megamodel through the execution of *global operation instances*. However the corresponding changes have not to violate the integrity of the megamodel which have to stay consistent.

C. The megamodel management

The first step in creating any composite service is to locate the service components (or tools) that provide the functionality that is to be placed in the new service [19]. To facilitate this process in our approach, all *component models* must be stored in a component directory, namely the megamodel, which can be accessed and managed at runtime. Each *component model* should be named and typed, and has also to specify what information (services) it represents in its corresponding integration dimension [7] represented by the model. *Component models* should also provide a description of its provided *operations* (through *global operation models*), its required *operations* (from other components), as well as the *input* and *output* services of all of these operations. Figure 1 represents the metamodel of the megamodel with respect to such considerations. Hereafter we present its main concepts.

- ***Model*** : Specifies that an entity is a model;
- ***ModelLevel*** : Allows knowing if it is a terminal model or a reference model ;
- ***ModelType*** : Allows giving the type of a model, to know its metamodel ;
- ***ModelDimension*** : Allows knowing the information represented in the model (data, process, ...);
- ***GlobalOperation*** : Supports operation on component models as :
  - ***Load*** : Registers a *component model* in the *Megamodel* if it is not already registered.
  - ***Extract*** : Allows the extraction of *component models* from the *Megamodel*.
  - ***Refine***: Allows to represent the refinement operation, i.e. transforming a model in a given dimension into another ;
  - ***Change2Ref***: Transforms a model in a given metamodel to another.
  - ***Match***: Takes as input two models of the same kind and performs especially well for detecting the differences between two versions of a *component model*

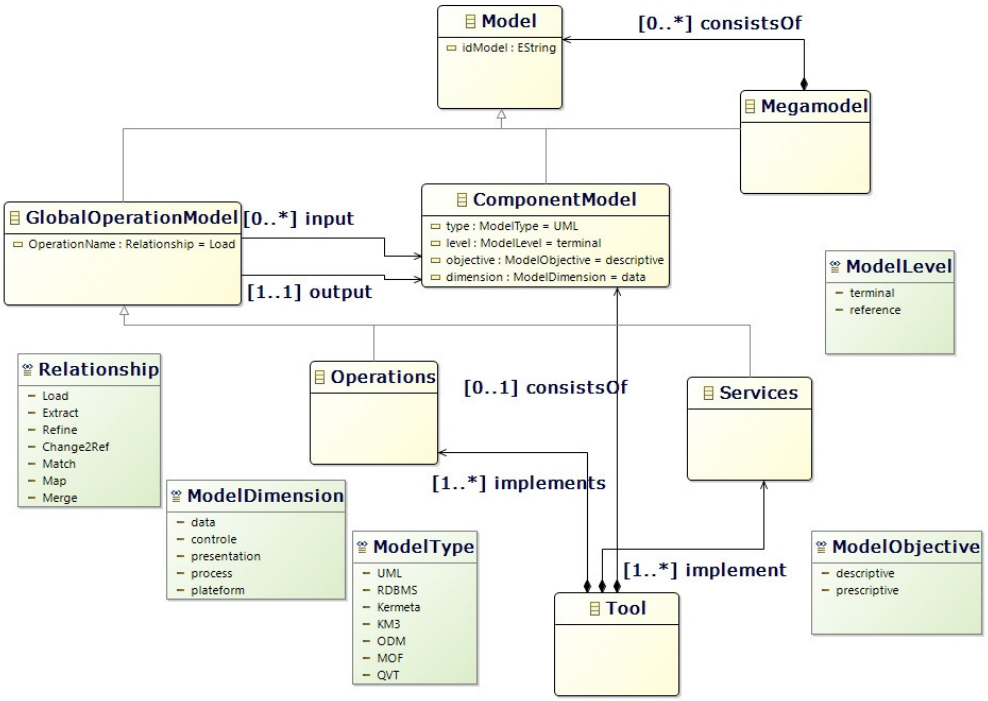


Fig. 1. Megamodel's Metamodel

- ***Map*** : Uncovers how two models "*correspond*" to each other. It takes two models as input and returns a morphism (which will be considered as a model) between them.
- ***Merge***: Merges two models into one based on a mapping (*Map*) between them.

V. Using Hoare's Axiomatics for Proving Megamodel's Consistency

A. Hoare Triples To Define Proof System of Axiomatic Semantics

In Hoare logic a program is considered a transformer of states. A state here represents the values of all the variables of a program. The execution of a program has the effect, if it ends, to transform an *initial state* to a *final state*. The specification of a program will be developing properties on these two states. *Axiomatic semantics* provides a style appropriate to the proof of programs. Thus, to prove a program, it just needed to specify the means of assertions written using *logical formulas* and then establish that the program meets its specification. That is to say that the specification and the program comply with the rules of an axiomatic semantics defining the valid executions of each statement in the source language. The technique used to prove an annotated program by assertions reduces it into a set of *logical formulas* called *verification conditions*, no longer refer to the instructions of the program. Prove a program thus reduces to *checking the validity* of logical formulas [9]. Properties of axiomatic semantics are expressed in general as expressions of Hoare logic: *{p}S{q}* where *p* and *q* are properties expressed in predicate logic, *p* supposed to be verified by memory before execution of the program *S*, and *q* to be checked after execution of *S* on the same machine that have checked *p*. Starting from the fact that one have to define a set of assertions that define as *pre-conditions* and *post-condition* rules of equilibrium of the system then the formalism of the Hoare logic is the most appropriate to represent the development of the possible states of the program, but also to make easier correction system. Considering the triple "*(x+1 > 0){ x := x + 1 }(x > 0)*". Such assertion means : *if " x+1>0"* is true before the execution of *" x := x+1 "*, *then* after his execution, the condition *" x>0"* becomes true. Or for the condition *" x>0"* is true after execution of *" x:=x+1"*, It is necessary that the condition *"x+1 > 0"* is verified, before performing the allocation.

B. Hoare's axiomatic semantics for proving megamodels consistency

As already said, the megamodel acts as an environment which is updated with each new *global operation execution*. A *global operation model* consists of a set of *pre-conditions* P, a *sequence of operations* Seq , and a set of *post-conditions* Q.

$$\{P\}Seq\{Q\}$$

The *pre-conditions* define the set of states of the system from which the *sequence of operations* can be invoked. Each *operation* is to be invoked in the order that it appears in the sequence, with the specified actual parameters and has possibly a *side effect* which corresponds to the effect of that operation

on other components. Indeed, each operation targets by default one or more *component models* or other *global operation models* in the megamodel. The *post-conditions* define the set of states that satisfy the desired result after executing the sequence of operations. The state of the megamodel is modeled by a set of assertions. And hence, for each *global operation models* to maintain this balance, it is necessary to know how its statements and flow effect affect correctness of models based on techniques available for proving programs consistency. To achieve, this we have defined as in our previous works [9], an *execution semantics* for each *global operation* in the megamodel. Execution semantics of a global operation in the megamodel denote its observable behavior (i.e. its effect on the state of the megamodel and components) as it is executed.

With this intention, it is enough that all the Hoare's triplets are valid and for this reason, it is necessary to formally define an execution semantic for each instance *Seq* of a *global operation model*.

VI. Example : Dynamic Tool Integration

A. Defining semantic execution for global operations

Suppose that we have a megamodel $\mathbf{M}$ which represents the overall structure of a software system $\mathbf{S}$ in terms of its constituent components (*component models*) and their interconnections (*global operation models*). Given that we have a new tool $\mathbf{T}$ to plug in the system $\mathbf{S}$. The tool $\mathbf{T}$ is then represented through multiple models, each one representing an integration dimension. $\mathbf{T}$ is said to be well integrated in $\mathbf{S}$ if for each dimension, the model representing a given dimension is well integrated with the model of $\mathbf{S}$ representing that dimension. To integrate two models we use the global operator called *Merge*. Our goal is not to say how to integrate two models (for that see [18]) but how the environment must react to the integration of new tools.

The algorithm (*see algorithm 1*) only gives the framework for implementing an environment supporting dynamic integration: it does not show how to implement an integration. When implementing an integration approach one has to recognize that the solution will never stay constant: new tools will be added, and, perhaps tools will be removed. This continuous change necessitates that the designer places emphasis on how the megamodel will evolve over time. Hoare's axiomatic semantics allow us to fix this problem by proposing formal safe and semantics execution for all the global operation instances in the megamodel. The execution semantics of a *global operation* is the effect of that operation on the state of the megamodel and will be defined by a rule.

In order to show how we use the axiomatic semantics for checking the megamodel's integrity, we consider an execution of the above algorithm. We assume that we have a system $\mathbf{S}$ which will be extended by plugging into it a tool $\mathbf{T}$. Given that the treatment is the same for all of the dimensions, we will consider in this example that $\mathbf{T}$ will be integrated to $\mathbf{S}$ according to the *data dimension*.

- $\mathbf{M}$ is the megamodel representing the system $\mathbf{S}$.

Algorithm 1 : Integrating new tool in the megamodel

1.**Input** : Software System $\mathbf{S}$, Tool $\mathbf{T}$, Megamodel $\mathbf{M}$
2.**Output**:$\mathbf{M}$ representing $\mathbf{S}$ in which $\mathbf{T}$ has been plugged
3.**BEGIN**
4. **For** *each dimension* d_i
5. **Begin**
6. **Let** ms_i *representing the dimension* d_i *of S*
7. $ms_i \leftarrow$ **Extract** (S, d_i)
. /* *Extract the model representing the dimension* d_i *in* $\mathbf{S}$ */
8. **Load** (M, ms_i)
. /* *Load the model representing the dimension* d_i *in* $\mathbf{M}$ */
9. **Let** mt_i *representing the dimension* d_i *of T*
10. $mt_i \leftarrow$ **Extract** (T, d_i)
. /* *Extract the model representing the dimension* d_i *in* $\mathbf{T}$ */
11. **Load** (M, mt_i)
. /* *Extract the model representing the dimension* d_i *in* $\mathbf{M}$ */
11. **If** (mt_i **AND** ms_i *have different metamodels*) **Then**
12. $mt_i \leftarrow$ **Change2Ref**(ms_i)
. /* *Transform a model in a given metamodel to another* */
13. **Let** map_i *be a morphism model defined as follows*
14. $map_i \leftarrow$ **Map**(ms_i, mt_i);
15. **Load** (M, map_i);
16. $ms_i \leftarrow$ **Merge**(ms_i, map_i, mt_i);
17. **Load** (M, ms_i);
18. **End**
19. **Return** M
20.**END.**

- ms_d is the model representing the data dimension of the system $\mathbf{S}$.
- mt_d is the model representing the data dimension of the system $\mathbf{T}$.
- map_d is the morphism (model) representing how ms_d and mt_d "*correspond*" to each other;

We will define an execution semantic for the *global operation* **Merge** enabling to integrate the tool T with the software S represented by the megamodel M according to their data dimension:

Operation : Merge

$merge_d \longleftarrow$ *Merge* (ms_d, map_d, mt_d)

Pre-conditions

$$P \Leftrightarrow \bigcap \begin{cases} P_1 & ::= & ms_d \in M \\ P_2 & ::= & level(ms_d, \text{"}\textit{terminal}\text{"}) \\ P_3 & ::= & type(ms_d, \text{"}\textit{UML}\text{"}) \\ P_4 & ::= & dimension(ms_d, \text{"}\textit{data}\text{"}) \\ P_5 & ::= & mt_d \in M \\ P_6 & ::= & level(mt_d, \text{"}\textit{terminal}\text{"}) \\ P_7 & ::= & type(mt_d, \text{"}\textit{UML}\text{"}) \\ P_8 & ::= & dimension(mt_d, \text{"}\textit{data}\text{"}) \\ P_9 & ::= & map_d \in M \\ P_{10} & ::= & merge_d \notin M \end{cases}$$

Post-conditions

$$Q \Leftrightarrow \bigcap \begin{cases} Q_1 & ::= & ms_d \in M \\ Q_2 & ::= & level(ms_d, \text{"}\textit{terminal}\text{"}) \\ Q_3 & ::= & type(ms_d, \text{"}\textit{UML}\text{"}) \\ Q_4 & ::= & dimension(ms_d, \text{"}\textit{data}\text{"}) \\ Q_5 & ::= & mt_d \in M \\ Q_6 & ::= & level(mt_d, \text{"}\textit{terminal}\text{"}) \\ Q_7 & ::= & type(mt_d, \text{"}\textit{UML}\text{"}) \\ Q_8 & ::= & dimension(mt_d, \text{"}\textit{data}\text{"}) \\ Q_9 & ::= & map_d \in M \\ Q_{10} & ::= & merge_d \in M \end{cases}$$

We have the triplet:

$$\{\text{P}\}\ \text{Merge}\ (ms_d,\ map_d,\ mt_d)\ \{\text{Q}\}$$

In the same manner, we define an execution semantics for each *global operation* defined in the megamodel. However, carrying out a *global operation* may require the execution of other *global operations* for standardizing the input models (ripple effect). This means that the output of a *global operation* may correspond to the inputs of another.

B. Inference rules for safe executions of global operations

The operation *Merge* takes as inputs two models representing the same dimension, in the contrary case, it would be necessary to call the *global operation model Refine*. Then if the two input models have not the same metamodel then we can also use the global operation *Change2Ref* enabling to transform a model specified in a given metamodel to another metamodel. Before applying the *Merge* operation, we have to earlier call the global operation *Map* with the two input models. Indeed the third parameter (*map*) to *Merge* is a morphism that describes elements of ms_d and mt_d that are equivalent and should be "*merged*" into a single element map_d in M. Once that all the models as inputs were described in the sound formalism, then one can apply the operation of merging. One thus realizes that it could exist a logical precedence between two or more *global operations* defined in the megamodel. In order to take into account these considerations we have to set up, in addition to axioms, a *deductive system* which permits the deduction of new theorem from one or more axioms or theorems already proved. A *rule of inference* takes the form "*If* $\vdash$ *X and* $\vdash$ *Y then* $\vdash$ *Z*", i.e. if assertions of the form *X* and *Y* have been proved as theorems, then *Z* also is thereby proved as a theorem.

For that we will use two rules of inference presented in [17], namely the *rule of consequence* and the *rule of composition*. After that, we present an example in which these two rules are applied.

(i) Rules of consequence :

If $\vdash$ *{P}S{R} and* $\vdash$ *R* $\supset$ *Q then* $\vdash$ *{P}S{Q}*
If $\vdash$ *{P}S{R} and* $\vdash$ *P* $\subset$ *Q then* $\vdash$ *{Q}S{R}*

These rules state that if the execution of a global operation ensures the truth of the assertion Q, then it also ensures the truth of every assertion logically implied by Q.

(ii) Rule of composition :

If $\vdash$ *{P}*S_1{Q_1} *and* $\vdash$ {Q_1}S_2{R} *then* $\vdash$ *{P}*(S_1; S_2){R}

This rule states that if the proven result of the first part of a *global operation* is identical to the pre-condition under which the second *global operation* produces its intended result, then both *global operations* will produce the intended result, provided that the pre-condition of the first *global operation* is satisfied. We will use the notation :

$$\frac{\vdash \{P\}S_1\{Q_1\} \ \vdash \{Q_1\}S_2\{R\}}{\vdash \{P\}(S_1; S_2)\{R\}}\ \textit{global operation}$$

We can then define the inference rule for the *Merge global operation* which enables us to carry out the integration of two models.

Considering the previous example. We have to describe an execution of *Merge* between the two models : ms_d and mt_d. However before applying the *Merge* Operation, it must be necessary to invoke at first the Operations : *Change2Ref* and *Map*.

Operation : *Change2Ref*
$mt_d \longleftarrow$ Change2Ref(ms_d)
Pre-conditions

$$P \Leftrightarrow \bigcap \begin{cases} P_1 & ::= & mt_d \in M \\ P_2 & ::= & ms_d \in M \\ P_3 & ::= & type(ms_d)! = type(mt_d) \\ P_4 & ::= & dimension(ms_d) == dimension(mt_d) \end{cases}$$

Post-conditions

$$Q \Leftrightarrow \bigcap \begin{cases} Q_1 & ::= & mt_d \in M \\ Q_2 & ::= & ms_d \in M \\ Q_3 & ::= & type(ms_d) == type(mt_d) \\ Q_4 & ::= & dimension(ms_d) == dimension(mt_d) \end{cases}$$

We have : $A_1 \longleftarrow$ *{P}Change2Ref{Q} (i)*

Operation : *Map*
$map_d \longleftarrow$ Map(ms_d, mt_d)
Pre-conditions

$$M \Leftrightarrow \bigcap \begin{cases} M_1 & ::= & mt_d \in M \\ M_2 & ::= & ms_d \in M \\ M_3 & ::= & map_d \notin M \\ M_4 & ::= & type(ms_d) == type(mt_d) \\ M_5 & ::= & dimension(ms_d) == dimension(mt_d) \end{cases}$$

Post-conditions

$$N \Leftrightarrow \bigcap \begin{cases} N_1 & ::= & mt_d \in M \\ N_2 & ::= & ms_d \in M \\ N_3 & ::= & map_d \in M \\ N_4 & ::= & type(ms_d) == type(mt_d) \\ N_5 & ::= & dimension(ms_d) == dimension(mt_d) \end{cases}$$

We have : $A_2 \longleftarrow$ *{M}Map{N} (ii)*

Using the **rules of consequence** on (i) and (ii) we obtain :

$\vdash$ *{P}Change2Ref{R} and* $\vdash$ *R* $\supset$ *M then* $\vdash$ *{P}MAP{N}*

We have : $A_3 \longleftarrow$ *{P}Map{N} (iii)*

Operation : *Merge*
$merge_d \longleftarrow$ Merge(ms_d, map_d, mt_d)
Pre-conditions

$$I \Leftrightarrow \bigcap \begin{cases} I_1 & ::= & mt_d \in M \\ I_2 & ::= & ms_d \in M \\ I_3 & ::= & map_d \in M \\ I_4 & ::= & merge_d \notin M \\ I_5 & ::= & type(ms_d) == type(mt_d) \\ I_6 & ::= & dimension(ms_d) == dimension(mt_d) \end{cases}$$

Post-conditions

$$R \Leftrightarrow \bigcap \begin{cases} R_1 & ::= & mt_d \in M \\ R_2 & ::= & ms_d \in M \\ R_3 & ::= & map_d \in M \\ R_4 & ::= & merge_d \in M \\ R_5 & ::= & type(ms_d) == type(mt_d) \\ R_6 & ::= & dimension(ms_d) == dimension(mt_d) \end{cases}$$

We have : $A_4 \longleftarrow \{I\}Merge\{R\}$ *(iv)*

Using the **rules of consequence** on (iii) and (iv) we obtain :

$\vdash$ *{P}Map{N} and* $\vdash I \supset N$ *then* $\vdash$ *{P}Merge{R}*

We have : $A_5 \longleftarrow$ *{P}Merge{R} (v)*

More formally, using the **rules of composition** we obtain :

$$\frac{\vdash\{P\}Change2Ref\{Q\} \vdash\{M\}Map\{N\} \vdash\{I\}Merge\{R\}}{\vdash\{P\}Merge\{R\}} \textit{Merge}$$

VII. Conclusion And Future Works

In this paper, we have presented a model-based environment of dynamic tool integration based on a MDE vision. We use a megamodel in order to manage dynamic changes in the software architecture. For that we have considered the megamodel as being a program and accordingly we have proposed an approach (namely Hoare's axiomatic semantics) already used for proving program's correctness which enables to keep the megamodel consistent.

As we look at future works, we will certainly be looking to set up Domain-Specific Modeling Language (DSML) which enables us to automatize the management of the megamodels and whose instructions will be the global operation instances. Therefore, each instruction, which consists of global operation execution, is likely to modify the megamodel structure and semantics. That is why such a DSML should also integrate an inference engine in order to provide a mean to reason about the elements of the megamodels. This is the inference engine that will be used to check the megamodels integrity and its elements through the validation of Hoare's triplets.

Acknowledgment

The research in this paper is supported by the *Centre d'excellence African en Mathématiques, Informatique et TIC (CEA-MITIC)*. We gratefully acknowledge the CEA-MITIC for financial support in that paper.

References

[1] Bezivin, J., Jouault, F., And Valduriez, P. *On the need for megamodels.* In Proceedings of the OOPSLA/GPCE: Best Practices for Model-Driven Software Development workshop, 19th Annual ACM Conference on Object Oriented Programming, Systems, Languages, and Applications.(2004, October).

[2] Iovino, L., Pierantonio, A., And Malavolta, I. *On the Impact Significance of Metamodel Evolution in MDE.* Journal of Object Technology, 11(3), 3-1.(2012).

[3] Bezivin ,J., Jouault, F., Rosenthal, P. And P. Valduriez. *Modeling in the Large and Modeling in the Small.*, In Proc. of European MDA Workshops: Foundations and Applications (MDAFA'05). LNCS 3599/2005, pp. 33-46. Springer, 2005.

[4] Kling, W., Jouault, F., Wagelaar, D., Brambilla, M., And Cabot, J. *MoScript: A DSL for querying and manipulating model repositories.* In Software Language Engineering (pp. 180-200). Springer Berlin Heidelberg.(2011).

[5] Seibel A., Neumann S., Giese H. *Dynamic Hierarchical Mega Models: Comprehensive Traceability and its Efficient Maintenance.* , Software and System Modeling 9(4):493-528, 2009.

[6] Vignaga A. , Jouault F., Bastarrica M. C., Bruneliere H. *Typing in Model Management.* , In R. F. Paige, editor, Second International Conference on Model Transformation. Theory and Practice of Model Transformations, volume 5563 of Lecture Notes in Computer Science, pages 197-212. Springer, 2009.

[7] Wasserman, A. I. *Tool integration in software engineering environments.* In Software Engineering Environments (pp. 137-149). Springer Berlin Heidelberg.(1990).

[8] Thomas, I and Nejmeh, B. A. *Definitions of tool integration for environments.* , IEEE Software, 9(2):29-35, March 1992.

[9] Bousso, M., Sall, O., Thiam, M., Lo, M., Toure, E. H. B. *Ontology Change Estimation Based on Axiomatic Semantic and Entropy Measure.* In Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on (pp. 458-465). (2012, November). IEEE.

[10] Fujii, K., and Suda, T. *Dynamic service composition using semantic information.* In Proceedings of the 2nd international conference on Service oriented computing (pp. 39-48). ACM.(2004, November).

[11] Mens, T., Mens, K., And Wuyts, R. *On the use of declarative meta programming for managing architectural software evolution.* In Proceedings of the ECOOP' 2000 Workshop on Object-Oriented Architectural Evolution, (2000, June).

[12] P. Atzeni, P. Cappellari, P. Bernstein *A multilevel dictionary for model management* , Int. Conf. on Conceptual Modeling (ER), Klagenfurt, Nov. 2005.

[13] T. Reiter, E. Kapsammer, W. Retschitzegger, W. Schwinger *Model Integration Through Mega Operations*, Workshop on Model-driven Web Engineering (MDWE), Sydney, July 2005

[14] G. Straw et al. *Model Composition Directives*, 7th UML Conference, Lisbon, 2004.

[15] Vogel, T., Giese, H. *A language for feedback loops in self-adaptive systems: Executable runtime megamodels.* (2012, June). In Proceedings of the 7th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (pp. 129-138). IEEE Press.

[16] Song, H., Huang, G., Chauvel, F., Sun, Y.: *Applying MDE Tools at Runtime: Experiments upon Runtime Models.* In: Models@run.time 10. CEUR-WS.org, vol. 641, pp. 25-36 (2010)

[17] Hoare, C.A.R., *An Axiomatic Basis for Computer Programming.* Communications of the ACM 12, 10, 576-583, October 1969.

[18] Pottinger, R. A., Bernstein, P. A. *Merging models based on given correspondences.* In Proceedings of the 29th international conference on Very large data bases-Volume 29 (pp. 862-873). VLDB Endowment (2003, September).

[19] Mennie, D., Pagurek, B. *An architecture to support dynamic composition of service components.* Systems and Computer Engineering. Carleton University, Canada. (2000)

[20] Blair, G., Bencomo, N., France, R. B. *Models@ run. time.* Computer, 42(10), 22-27.(2009).

[21] Salay, R., Mylopoulos, J., Easterbrook, S. *Using macromodels to manage collections of related models.* In Advanced Information Systems Engineering (pp. 141-155). Springer Berlin Heidelberg. (2009, June).

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 219–224
DOI: 10.15439/2016F180
ACSIS, Vol. 9. ISSN 2300-5963

Competences in the knowledge-based economy

Jolanta Sala
Powiślański College,
ul. 11Listopada 13, 82-500
Kwidzyn, Poland
Email: jolasala@interia.pl

Halina Tańska
University of Warmia and Mazury
in Olsztyn, ul. Słoneczna 54,
10-689 Olsztyn, Poland
Email: tanska@uwm.edu.pl

Abstract— **The article discusses the causes of the difficulties and contradictions associated with the management information systems referring to the competences of knowledge workers in organizations. As far as entities are concerned, the reasons are mainly the effects of global phenomena of socio-economic life. Unfortunately, with the increase in the declared need for competences in the knowledge-based economy, there are a growing number of knowledge workers without jobs and livelihood. Social inequalities are rising. At the same time, the productivity in enterprises is increasing, the profits are growing, and the wages are declining. Knowledge workers without stable jobs are the new dangerous class "precariat" of enormous magnitude. There are very serious threats which require solutions for the management of information systems referring to the competences of knowledge workers at all levels of socio-economic life.**

I. Introduction

The authors of this paper face difficulties with the identification [1], measurement, assessment and analysis of competences of knowledge workers on a daily basis, which universities struggle with. Competences of knowledge workers are also a challenge in other organizations, in particular in the commercial companies. A competence in the knowledge-based economy is a complex category based on three attributes: knowledge, skills and approaches adjusted to the situation (capabilities, personal, psychosocial and cognitive predispositions). A competence is often equated with qualifications. The problem of information systems in this context, has been signalled methodically by J. Oleński in 2000 [2], who has indicated the cause of standardization deficiencies in the system of national accounts SNA 93 and ESA 95. J. Oleński identified the paradoxes of information systems, which the authors of this study verified in their studies, and in particular since 2010 in enterprises in the Pomeranian Province. The aim of this study is to identify the most serious consequences of underdevelopment of information systems (ISM) in the context of competence.

.

Managing knowledge workers is not an easy task, and the topic is still present in some economic and social institutions, local and regional communities, and also on a national and global scale. The paper presents selected aspects of competences in the knowledge-based economy from the perspective of knowledge workers, who have become a part of the emerging social group called a precariat [3]. The precariat is a social group newly forming on the global scale, whose characteristics are so different from past experience, that not only in theory, but also in practice (managers and politicians) there is no understanding of its essence and size. The precariat create new requirements for IT systems referring to managing knowledge workers. In this paper, the authors synthetically present assumptions and main conclusions of the study, whose aim was to demonstrate on a small group the phenomenon of the lost human capital potential, not only in enterprises but also in the region due to the underdevelopment of information systems in the context of professional competence.

II. Economic and social prosperity

The essence of the industrial economy was land, capital, labour (work) and its harmonious development was provided for example by reorganization of rules connected with working conditions. Labour is understood as human efforts made to produce the good along with the price, which is a wage. In the classical economics of A. Smith (1776) two other factors are land and natural resources owned by people (a price of land is a land rent) and capital understood as goods produced previously by people and used to the production of other goods – such as buildings, tools and machinery (the price of the capital is interest rate). The contemporary world entered the postindustrial economy based on globalization and information society - often called the knowledge-based economy - with the exposed factor of mental work (white-collar workers) focused on information and knowledge, which is shown in the Figure 1.

This work was not supported by any organization

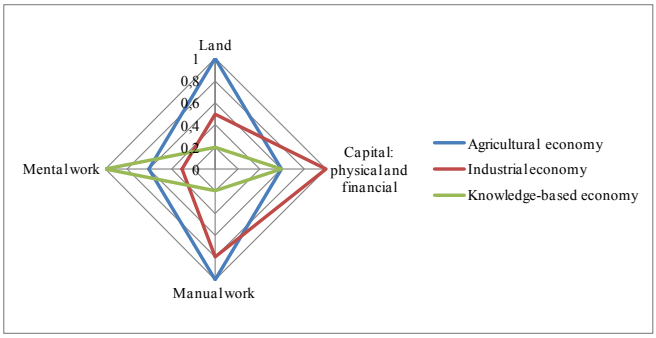


Fig. 1 The comparative interpretation of agricultural, industrial and knowledge-based economies [own research]

There have been significant changes in existing rules connected with work management and organization. For instance, enormous expectations were linked to the flexible forms of employment, which consequently required elaboration of the safety rules - flexicurity. Studies conducted by economists and sociologists, including J. Stiglitz [4], M. Castells [5], J. Dijk [6], revealed many negative implications of the existing changes. M. Castells formulated a diagnosis of consequences of ICT technologies as ten processes [7, pp. 60-64]: rising inequalities and social exclusion all over the world (process 1), appropriation of the wealth generated by collective effort (process 2), four processes connected with the relationships of distribution/consumption, i.e. inequality, polarization, poverty, misery, as well as four processes connected with relationships of production, i.e. individualization of work, over-exploitation of workers, social exclusion, perverse integration.

The negative implications referring to the impact of ICT technologies identified by M. Castells have been confirmed by G. Standings' conclusions on changes taking place in the social class structure. Standing has identified a new social group 'precariat', presented in Figure 2.

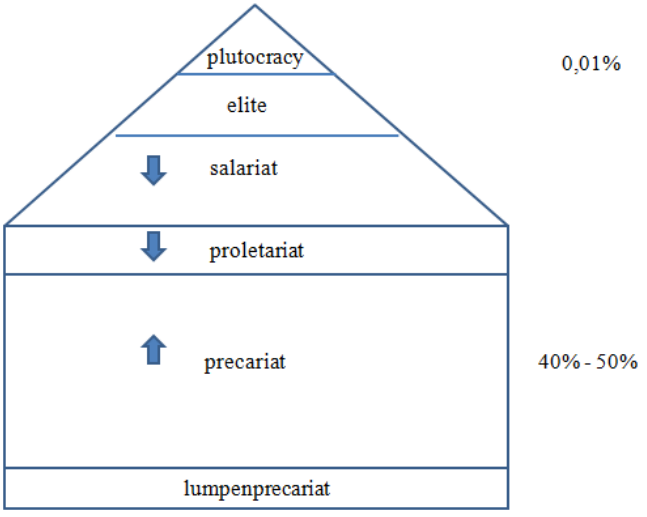


Fig. 2 The new hierarchy of class structure of the adult population of society emerging as a result of globalization according to G. Standing [8]

According to G. Standing [8], at the top of the hierarchy there is a plutocracy, and below are elite, a salariat, a proletariat, a precariat and a lumpenprecariat. Plutocracy are 0,01% of the richest people, hundreds of people, who own the capital worth approximately 20-40 billion USD, and at the top of this group there are people owning 80-90 billion USD! They have great power, mightiness, they can have impact on media, sponsor politics and fund political parties. Elite can be distinguished by huge capital gains and whereas Salariat is characterised by stable employment, high salaries, access to capital gains, variour benefits, pension schemes, etc. Proletariat is disappearing all over the world; Precariat is a newly emerging social group. Lumpenprekariat are people expelled from society, suffering from different types of mental disorders, homeless, isolated.

According to G. Standing's estimates [9], the precariat accounts for above 40% of adult population of societies in the developed countries, and according to studies conducted in Japan – over 50% are a group of precarians, i.e. people deprived of stable employment and doomed to have a job below their level of educational qualifications. The social group 'precariat' is defined in three dimensions [3]: relationships of production, relationships of distribution, political relationships. The precariat has been spreading and growing at a faster pace since the crisis in the year 2008 as a result of the policy of austerity measures and budget cuts. Employers take the opportunity to reduce costs through a deterioration of working conditions. Sadly, knowledge workers account for a large part of the precariat, which is not a good sign of development of the knowledge-based economy [10].

The authors [11] have attempted to identify the prosperity in context of labour and tools, crucial for fulfilling the task, which has been recently noted by Poles seeking the job. The authors have distinguished seven phenomena of social and economic life in Poland and with the help of those phenomena they showed the main weak points of work processes. The authors have exposed the neglects connected with the relation of work and ICT tools as the vital foundation of building the prosperity in XXI century.

On the one hand worldwide and Polish experiences still bring negative outcomes, but on the other hand there is a search for the transformation of society and the transformation of organizations with „human face" and reflecting capabilities and the potential of new technologies. The achievements of Scandinavian countries, such as Finland and Sweden in particular, are often presented as the example of economic and social prosperity.

III. COMPETENCES FRAMEWORK AND QUALIFICATIONS FRAMEWORK

In times of the knowledge-based economy more than 135 countries have decided to introduce qualifications framework (QF), also the EU countries have been involved in the EQF. Qualifications framework systematises knowledge, skills,

approaches and is based on two dimensions of the competences: general education and vocational education. Polish qualifications framework (PQF) comprises eight levels. The graph in the Figure 3 presents a synthetic interpretation of relationships between both dimensions. Because of the substantial collapse of formal vocational education in Poland one can say, that general education has formal character (the Y-axis), and vocational education – non-formal and informal (the X-axis). Although the graph was constructed on the basis of studies conducted in Poland, its interpretation is universal.

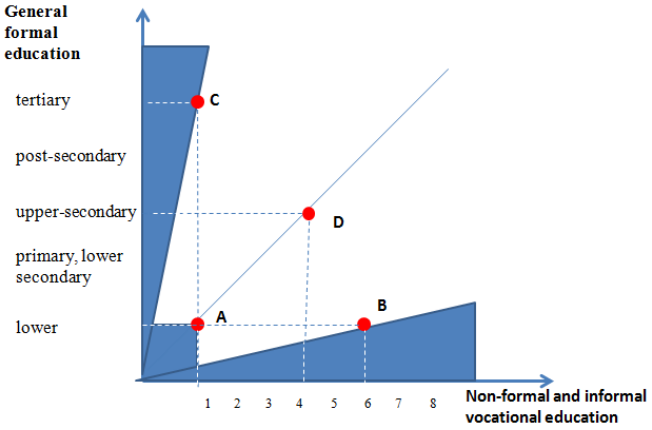


Fig. 3 Relationships of identified cases in the Polish education system [own research]

The point A in Figure 3 (and the area it belongs to) describes a relatively small group of workers, just like the point B. However, a large group of workers and potential workers, particularly young ones, are represented by the point C. Obviously, the point D is ideal in terms of effectiveness, balancing theoretical and practical potential. It can be assumed, that while moving upwards a diagonal, scribed by points AD, the importance (value) of a worker is increasing for the knowledge-based economy.

The analyses, which are being conducted by the Polish Ministry of Digital Affairs, have led to the elaboration of a very useful model of digital competences, which are the central attribute of key competences for the knowledge-based economy [12]. According to the EU [13], key competences are those, which all people need for fulfilment and personal development, being an active citizen, social integration and employment. The following eight key competences have been established: (1) communication in the mother tongue; (2) communication in foreign languages; (3) mathematical competence and basic competences in science and technology; (4) digital competence; (5) learning to learn; (6) social and civic competences; (7) sense of initiative and entrepreneurship; (8) cultural awareness and expression. The relational model [12] was identified on the three levels of digital competences: IT, information and functional. As a result, the framework catalogue of digital competences was prepared, which might be the reference point for activities aimed at providing and enhancing digital competences. An important assumption of the catalogue is the connection between the digital competences and the users' needs and the benefits that they may gain in key areas of life.

IV. Transformation of Organizations

In case of a university – which is an interdisciplinary and multifunctional organization – management of knowledge workers is often passive, random or discretionary. Sadly, it does not foster creativity, and undoubtedly is dominated (overwhelmed) by bureaucratic rules. A lot of expectations are based on internal quality management systems in education and other rules of workers assessment. There are various concepts regarding management of knowledge workers in companies, who see their opportunity in transformation and gaining competitive advantage for example by taking part in digital industrial revolution [14], [15] or affirmation of idea "Industry 4.0".

A proposal of the approach to the individual evaluation of a knowledge worker was presented in the paper [16] and is worth considering. It is focused on developing a profile of an engineer (automation service engineer, computer scientist or ITmatician) in the context of eight dimensions of quality: effectiveness, functionality, reliability, durability, compliance with standards, "supportability", aesthetics, "perceived quality". Although assumptions of the approach by [17] seem to be controversial, but, undoubtedly, the direction of adjusting quality of job carried out by knowledge worker to needs and expectations of a customer is the proper one. The core of the approach is to deliver a tool supporting innovative projects for both parts, employers and contractors, in the form of a „image of quality" presented in the Figure 4.

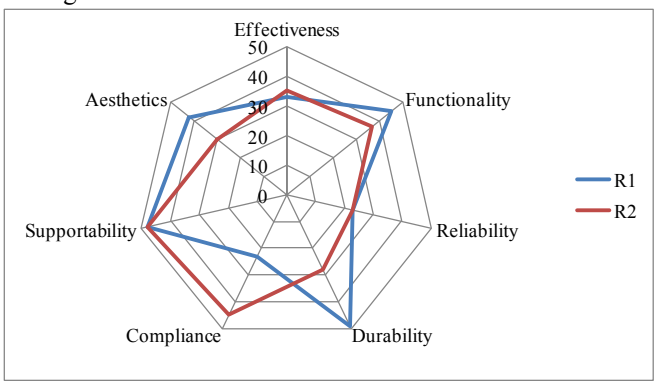


Fig. 4 „Image of quality" of knowledge workers competences from selected vocational groups [16]

The graph in the Figure 4 shows seven dimensions of quality, which are universal for the area of knowledge represented by two sample solutions. Analyzing the eighth dimension - „perceived quality" - is recommended individually for each solution (project) or even for each person taking part in the project. With this tool a contractor of the solution R1 and/or R2 has the support before making

irresponsible promises (mainly due to the possession of an „image of quality" regarding competences of knowledge workers acting as sellers and performers). Similarly for employers, taking a look at competences of companies (through knowledge workers representing them) who offer both solutions is a useful „image of quality" for a planned innovation.

The authors considered this tool aspect, defining (identifying) knowledge cage for project teams [18] and trying to analyze a "pig in a poke" syndrome ("cat in the sack" syndrome) [19]. Not only the knowledge-based economy, but also information society needs this type of tools, although their practical usage is a long-term process. A good example are difficulties connected with evaluating intellectual capital of companies and applying to them generally accepted accounting principles, which is described with determination in publications written by i.a. L. Niemczyk [20].

The identification and analysis of information gap does not lose its instrumental up-to-date character, including competence gaps, which the authors focused on in studies and many publications, for example [21]. However, due to dynamic changes and development of science and technology in social and economic life there is non-trivial problem regarding the model for identifying gaps, and the risk of uncritical trust in models. Many companies and a lot of their workers recorded losses in a long-term perspective, which had been caused by affirmation of flexible forms of employment providing short-term gains. Life begins to be consumed by the work/activity.

Attempts aimed at moderating the knowledge-based economy in the EU have led to the rule of forming "intelligent specializations" for countries and regions.

V. The Working Conditions and Transformation in Poland

Transformation in Poland after 1989 was also associated with the necessity of the creation of public employment services, which have matured to the role of socio-economic, along with the formation of the labor market. Parallel changes were subject to information systems for monitoring and evaluation of the labor market. Therefore it is worth noting synthetically the current state of information systems of Polish public employment services, although it would be interesting to apply the possibilities of a conversion method for the evaluation [22] and assess the quality of e-government portals [23]. The authors focused on information as an economic resource, as a common good and as an economy infrastructure in line with the aspects approach of J. Oleński [2]. Despite the development of a new systems of national accounts SNA 2008 and ESA 2010 it is still not possible to classify the resources important for the knowledge-based economy.

The public information system for Polish employment services looks notably modest, although the progress is undoubtedly large. They are integrated and standardized different information systems, run by various entities, providing the Internet vortal [24]. From the perspective of the labor market the portal provides five "databases", including the classification of professions and specialties, and standards / professional competencies base and modular training programs. In addition, the portal also provides "statistics and analysis", but unfortunately only in two areas: the "registered unemployment" and "employment of foreigners in Poland". During the integration of information systems in the vortal there were ignored reports on "competition deficit and surplus" available only on the website of the Ministry of Family, Labour and Social Policy [25].

As to "strategies and programming documents" on the vortal there is a strategy for the development of human capital, which is a copy of the strategy "Europe 2020". Among the attachments three are two recent documents: "The project of the National Action Plan for Employment in the period 2015-2017" and " Program and organizational assumptions for preparation of the National Action Plan for Employment in the years 2015-2017". All program goals are grouped in the two following priority areas: increasing the efficiency of the management of the labour market in order to support the growth of employment and increase of adaptability of the labour market. Undoubtedly, this information system would need the risk analysis in accordance with the conclusions of [26], as well as the use of idea electronic communities published in [27].

The direction of changes which was indicated by G. Standing, resulting in the depreciation of knowledge workers who have been respected and successful in professional life so far, is also confirmed by consequences of restructuring of companies in Pomeranian (pomorskie) voivodeship at the turn of 2014 and 2015 year. The authors have conducted research based on 241 people (20F/221M), who experienced unexpected personal trauma. Virtually overnight, mainly engineering and technical workers found out, that their knowledge, skills and experience were worthless, regardless of their formal education (19% of whom were experienced employees with university degrees), which is shown at the graph in the Figure 5 (also regardless of their non-formal and informal education, qualifications, played roles and occupied positions – over 60% of them were highly qualified technicians).

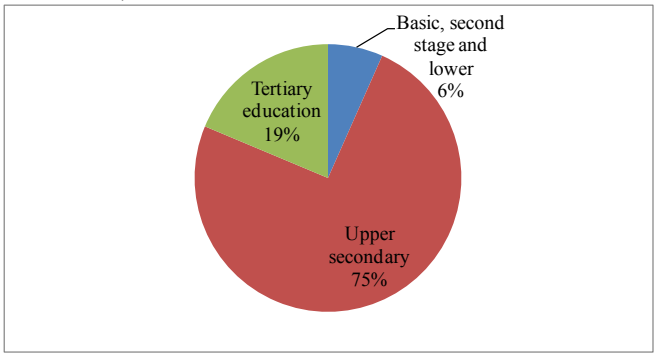


Fig. 5 A structure of the researched group by formal education [own research]

Due to intervention of state institutions, people from the researched group could have tried to become entrepreneurs, starting their own companies with the financial support (grants) or acquire new qualifications, taking part in free trainings. Because of the transformation of companies (92% of big enterprices) a group of the precariat has increased on the Pomeranian labour market, which does not mean that there has been an increase in the number of unemployed. The companies, in line with their own decisions, has lost their intellectual capital. The personal situation of employees made redundant for reasons attributable to the employers has deteriorated drastically, despite the state supportive intervention. The analyzed group of 241 people accounts for a part of wider studies which have been conducted by the authors, involving 2094 people, since the year 2010, and a phenomenon of the fast growing precariat is widespread. In the context of this fact, the knowledge-based economy and its main assumptions seem to be a utopia and require a fundamental revision.

Similar state intervention as in the Pomeranian Province were undertaken during the years 2007-2014 and are planned in the years 2014-2020 also in other provinces. During the study it was verified that this type of government intervention to increase the efficiency and adaptability of the labor market. However, emerging information resources were used mainly for the settlement of this business. Contractor of intervention was obliged to collect data in three information systems: PEFS 2007, monitoring of indicators and the Registry "employability".

The PEFS 2007 served in the years 2007 to 2014 in the records of the persons (people) receiving support from national and European funds (ESF). Monitoring of indicators was processed normally on a quarterly basis and served as quantitative reporting in 7 layouts (input-output, by employers, by age, by education and other specific indicators). Records of "employability" was used as a quantitative monitoring of two states employment of persons covered by the support i.e. at 3 months and 6 months after termination of support. Information resources of these systems were very dispersed and were only operating instruments although their potential was much greater.

There was probably an effort taken for the integration of the functionality of those three information systems into a single system SL2014, which the authors plan to verify in the terms of the range of information and support of decision-making processes in the context of professional competencies, the phenomenon of loss of human capital, and tackling the problem of the depreciation of precariat. Unfortunately, the research hypothesis is not optimistic, which is not a good anticipation for the development of knowledge-based economy. Development of the knowledge-based economy cannot be based only on companies but also on the added value of various state intervention, including similar to those covered by the research.

VI. Conclusion

The knowledge-based economy in practice deviates significantly from the theoretical assumptions. Transformations taking place in organizations and countries have led to negative economic and social consequences.

What the studies confirm is that the level of economic and social prosperity has decreased significantly in the developed countries. The adverse effects of existing changes are heavily felt by educated people, who, in principle, were to be the pillar of the knowledge-based economy (Figure 1). The mental work - as the most crucial production factor - has relationships in production highlights the tendencies depreciating possibilities enabling to keep the pace of development for science, technologies and culture, e.g. the individualization of work. As a result, the economic and social situation of the new social class ‘precariat’ is deteriorating (Figure 2). According to G. Standing, for the first time in history of economics, the productivity is growing, capital gains are increasing, but salaries are decreasing.

For the organizations who are in the period of transition to the knowledge-based economy, the quality management is still a useful tool. The instrumental potential is still associated with the improvement of systems and information processes, i.e. information gap, competence gap, information asymmetry, etc. Thus, the information economics indicates directions in implementations of information systems supported by ICT technologies. Applications of ICT technologies, assisting various areas of managing organizations in transitions, are still pioneers and creators of directions of changes, strengthening the position of ICT industry. The industry does well with commercialization and marketing of its offers starting from the highest political levels (the latest examples are the fourth digital-industrial revolution and Internet of Things along with standards). Developing models and applying benchmarking is seen as crucial. There is a growing need for instruments of accounting, which evaluate intellectual capital.

With the alarming state of the knowledge-based economy in the context of precariat and complex situation on the labour market, companies have intensified their marketing efforts and raised interests in the applications enabling human resources management. The need „lack of well qualified specialists” was identified as well as the set of solutions supported by ICT applications, including paying more attention to the development of already employed workers and the better matching workers to the positions. Thus, „the competence mismatch, which results in the low effectiveness of workers” [28], has become a marketing argument for the implementation of advanced IT tools as the only antidote for large and medium enterprises.

It should be underlined that the issue of knowledge workers’ productivity is not at the core of the current global collapse of the knowledge-based economy. If direct employers and indirect employers (state institutions) do not

react in relatively fast and effective way to difficult, sometimes even dramatic economic and social situations of knowledge workers - who are pushed to the class of precariat, the development of the knowledge-based economy will be obstructed because of the depreciation of high competences, education and lifelong learning. Currently, a dramatic waste of the intellectual potential is continuing and is growing on the global scale. Managers and politicians are not able to manage intellectual capital and social capital, putting themselves first, while the scale and pace of enrichment is unimaginable. As far as Poland is concerned, the situation and conclusions are similar, although in many ways they reflect the specificity of the country.

I. References

[1] I. Pawełoszek, "Semantic Organization of Information Resources for Supporting the Work of Academic Staff," Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1235–1241, ACSIS, Vol. 2, DOI: 10.15439/2014F320

[2] J. Oleński, Elementy ekonomiki informacji, Katedra Informatyki Gospodarczej i Analiz Ekonomicznych, Uniwersytet Warszawski, Warszawa 2000.

[3] G. Standing, The Precariat. The New Dangerous Class, Bloomsbury Academic, London 2011

[4] J. E. Stiglitz, Szalone lata dziewięćdziesiąte, Wydawnictwo PWN, Warszawa 2006

[5] M. Castells, Społeczeństwo w sieci, Wydawnictwo Naukowe PWN, Warszawa 2008.

[6] J. Dijk, Społeczne aspekty nowych mediów. Analiza społeczeństwa sieci, Wydawnictwo Naukowe PWN, Warszawa 2010.

[7] M. Castells, Koniec tysiąclecia, Wydawnictwo Naukowe PWN, Warszawa 2009.

[8] G. Standing, "The Precariat and Class Struggle," Revista Crítica de Ciências Sociais 103, May 2014, p.9-14.

[9] E. Koszowska, „Bezwarunkowy dochód podstawowy. Prof. Guy Standing: każdemu się należy," WP Opinie [on-line]. Grupa Wirtualna Polska S.A., 2015-10-21

[10] J. E. Stiglitz, Cena nierówności. W jaki sposób dzisiejsze podziały społeczne zagrażają naszej przyszłości, Wydawnictwo Krytyki Politycznej, Warszawa 2015.

[11] J. Sala and H. Tańska, „Dobrobyt w kontekście pracy i narzędzi pracy," in Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2015.

[12] J. Jasiewicz, M. Filiciak, A. Mierzecka, K. Śliwowski, A. Klimczuk, M. Kisilowska, A. Tarkowski and J. Zadrożny, Ramowy katalog kompetencji, MC, Warszawa 2015, https://mc.gov.pl/files/ramowy-katalog-kompetencji-cyfrowych.pdf

[13] Zalecenie Parlamentu Europejskiego i Rady Unii Europejskiej nr 2006/962/WE z dnia 18 grudnia 2006 w sprawie kompetencji kluczowych w procesie uczenia się przez całe życie [Dz.U.L394 z 30.12.2006].

[14] A. Gąsiorek, „Cyfrowa rewolucja przemysłowa szansą na transformację i zdobycie przewagi konkurencyjnej," Control Engineering nr 2 (118) 2016, Wydawnictwo Trade Media International, s. 127.

[15] M. Łopaciński and M. Matejczyk, „Transformacja organizacji – jak ją przygotować aby miała szansę powodzenia," Biznes i produkcja, nr 13 (2/2015), Wydawca ASTOR, s. 6-9.

[16] J. Gracel, „Jakość inżyniera ...," Biznes i produkcja, nr 11 (2/2014), Wydawca ASTOR, s. 42-43

[17] D. A. Garvin, "Competing on the Eight Dimensions of Quality," Harvard Business Review, November-December 1987, pp. 101-109.

[18] J. Sala and H. Tańska, "The cage of knowledge in process of information systems development," in Computer Systems Engineering Theory & applications, K. J. Burnham, L. Koszalka, Wrocław 2005, ISBN 83-911675-8-5, s. 157-167.

[19] J. Sala and H. Tańska, „Syndrom „kota w worku" w społeczeństwie informacyjnym," in Społeczeństwo informacyjne w świecie rzeczywistym i wirtualnym, A. Szewczyk (red.)], Zeszyty Naukowe Uniwersytetu Szczecińskiego, nr 656, Studia Informatica nr 28, Szczecin 2011, s. 455-465.

[20] L. Niemczyk, Kapitał intelektualny w księgach rachunkowych oraz sprawozdawczości przedsiębiorstwa, Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów 2015

[21] J. Sala and H. Tańska, "Tool dilemmas of innovation," Position papers of the 2014 Federated Conference on Computer Science and Information Systems pp. 265–269, ACSIS, Vol. 3, DOI: 10.15439/2014F484.

[22] W. Chmielarz and M. Zborowski, "The Application Of A Conversion Method In A Confrontational Pattern-Based Design Method Used For The Evaluation Of IT Systems," Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1227–1234, ACSIS, Vol. 2, DOI: 10.15439/2014F198.

[23] E. Ziemba, T. Papaj and D. Descours, „Assessing the quality of e-government portals – the Polish experience," Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1259–1267, ACSIS, Vol. 2, DOI: 10.15439/2014F121.

[24] http://psz.praca.gov.pl.

[25] http://www.mpips.gov.pl/analizy-i-raporty/raporty-sprawozdania/rynek-pracy/zawody-deficytowe-i-nadwyzkowe/

[26] E. Ziemba and I. Kolasa, „Risk factors framework for information systems projects in public organizations – Insight from Poland," Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1575–1583, ACSIS, Vol. 5, DOI: 10.15439/2015F110.

[27] C. Stepniak and T. Turek, „Levels of the Use of Electronic Communities in the Management of Regions," Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1551–1556, ACSIS, Vol. 5, DOI: 10.15439/2015F277.

[28] M. Pawlik, „Tendencje na polskim rynku teleinformatycznym w 2015," Inżynieria & utrzymanie ruchu, nr 1 (112), Wydawnictwo Trade Media International, 2016, s. 12-13.

11th 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

- 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

- **Abu-Shanab, Emad,** Yarmouk University, Jordan
- **Alghamdi, Saleh,** University of Sussex, United Kingdom
- **Bialas, Andrzej,** Institute of Innovative Technologies EMAG, Poland
- **Bicevska, Zane,** DIVI Grupa Ltd, Latvia
- **Chung, Tsungting,** Douliou Yunlin Uniwersytet, Taiwan
- **Czarnacka-Chrobot, Beata,** Warsaw School of Economics, Poland
- **DeLorenzo, Gary,** California University of Pennsylvania, United States
- **Dima, Ioan Constantin,** Valahia University of Targoviste, Romania
- **Duan, Yanqing,** University of Bedfordshire, United Kingdom
- **Dudycz, Helena,** Wrocław University of Economics, Poland
- **El Emary, Ibrahim,** King Abdulaziz Univetrsity, Saudi Arabia
- **Espinosa, Susana de Juana,** University of Alicante, Spain
- **Geri, Nitza,** The Open University of Israel, Israel
- **Grublješič, Tanja,** University of Ljubljana, Slovenia
- **Halawi, Leila,** Embry-Riddle Aeronautical University, United States
- **Jankowski, Jarosław,** West Pomeranian University of Technology in Szczecin, Poland
- **Jelonek, Dorota,** Czestochowa University of Technology, Poland
- **Kobyliński, Andrzej,** Warsaw School of Economics, Poland
- **Michalik, Krzysztof,** University of Economics in Katowice, Poland
- **Mullins, Roisin,** University of Wales Trinity Saint David, United Kingdom
- **Niedźwiedziński, Marian,** University of Lodz, Poland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland

- **Ozkan, Necmettin,** Turkiye Finans Participation Bank, Turkey
- **Pastuszak, Zbigniew,** Maria Curie-SKlodowska University, Poland
- **Ranjan, Jayanthi,** Institute of Management Technology in Ghaziabad, India
- **Ren, Kun,** Yale University, United States
- **Rizun, NIna,** Alfred Nobel University, Dnipropetrovs'k, Ukraine
- **Schroeder, Marcin,** Akita International University, Japan
- **Sikorski, Marcin,** Gdańsk University of Technology, Poland
- **Silber-Varod, Vered,** The Open University of Israel, Israel
- **Skovira, Robert,** Robert Morris University, United States
- **Sobczak, Andrzej,** Warsaw School of Economics, Poland
- **Surma, Jerzy,** Warsaw School of Economics, Poland
- **Świerczyńska-Kaczor, Urszula,** Jan Kochanowski University in Kielce, Poland
- **Symeonidis, Symeon,** Democritus University of Thrace, Greece
- **Szczerbicki, Edward,** University of Newcastle, Australia
- **Tarhini, Ali,** Brunel University London, United Kingdom
- **Travica, Bob,** University of Manitoba, Canada
- **Wachnik, Bartosz,** University of Technology in Warsaw, Poland
- **Wątróbski, Jarosław,** West Pomeranian University of Technology in Szczecin, Poland
- **Wielki, Janusz,** Opole University of Technology, Poland
- **Wolski, Waldemar,** University of Szczecin, Poland
- **Ziemba, Paweł,** The Jacob of Paradyż University of Applied Science, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 227–231
DOI: 10.15439/2016F591
ACSIS, Vol. 9. ISSN 2300-5963

Innovation In Energy: Dissemination Of Energy Culture Via Serious Gaming

Esra Çalışkan, Gülgün Kayakutlu , Vehbi Tufan Koç
Industrial Engineering Department,
Istanbul Technical University
Maçka Istanbul 34367, Turkey
Email: {caliskanesra, kayakutlu, koctu}@itu.edu.tr

***Abstract*—In today's world, the term serious game is becoming more and more popular since these games have a purpose of engaging the user and contributing to the achievement of a specific goal other than pure entertainment games. Serious gaming has high potential and power to transfer knowledge and educate people. Serious games are applied to a large spectrum of application areas ranging from military, government to healthcare. Also, researchers found out that serious gaming is a potential tool in order to enable young people to understand the complexity of sustainability and energy conservation topics and stimulate the energy saving attitudes and behaviors and arising energy conservation awareness in a fun way. This study purposes to disseminate energy culture via developing a serious game for university students. Thus, the paper focuses on serious gaming with a broad range of aspects including the need of serious gaming in various sectors and its benefit for behavior change and its effects on energy culture. While developing the game, the conceptual framework which is called as 'Design Pattern Library' is considered. Description and the flow of the game are determined and the game is played between students. Results obtained by this study will enlighten both researchers and game design experts in terms of disseminating energy culture via serious gaming.**

***Index Terms*—Serious Gaming, Energy Culture, Behavioral Change**

I. Introduction

IT IS known that the idea of playing a game has emerged since the ancient past and it is still on going as an integral part of today's world. The Dice game which is one of the oldest games discovered by humans is known to be set 3000 years ago in South Iran (as cited in Laamarti et al., 2014). Games are played according to a set of established rules, played out within a specific time frame and a place with a player or a group of players. Indeed, games have always been powerful to affect people's lives in many aspects and still games have huge power on people wherever in the world. It is known that games have power to bring people all together from 7 to 70 and socialize. It is claimed that games are even powerful in revealing and building people's characters (Michael et al., 2006) Most importantly, games have power to teach, to educate and to train. Since most games are claimed to have a primary purpose of entertainment, enjoyment and fun way to pass the time or interact with other players, serious games are distinguished with their serious purposes. IDATE pointed out three main principles of serious games: delivering a message, providing training and enabling the exchange of information (2010). These principles briefly show that serious games have a goal to transmit a message with a serious purpose, to improve the users' cognitive and/or motor skills via trainings and to provide an environment for flow of data between users. Besides, Knol and Vries supported that serious gaming is a potential tool in order to enable young people to understand the complexity of sustainability and energy conservation topics and stimulate the energy saving attitudes and behaviors and arising energy conservation awareness in a fun way (2011). Most importantly, these serious games are explicitly designed to encourage behavioral change. As this study aims to disseminate energy culture via serious gaming, the next part consists of literature review on serious gaming. Also, sample games that aim to change energy related behavior patterns are analyzed. The third part explains how the methodology works for developing a serious game. The game board and the flow of the game 'Enopoly' is given in the fourth part. The fifth part discusses the results of the game and gives suggestions for further studies. This study aims to contribute both researchers and game developers.

II. Literature Review

A. Serious Gaming in Various Sectors

There is a growing evidence that some people responds better to "playful" kind of formats of information feedback and this is why the serious gaming trend raises by using digital games for learning, simulation and demonstration (Wood et al., 2014). Inevitably, the serious game sector is showing significant growth in the medium term. According to IDATE's Serious Gaming Market and Data Report (2010), it is stated that serious gaming is an attractive sector and the players in the sector are grouped into 4 categories as seen on the figure below: Traditional software sector players, groups of investors, intermediate players and the targeted sectors.

The report emphasized on targeted sectors category are mainly consisted of healthcare, teaching and training, information communication promotion, defense and civil security. Indeed, there is a wider segmentation since serious gaming applications can be used in any sectors. To begin with healthcare sector, it is remarked that sector specific aspects are broad and serious games is widely used from marketing of medication brands to provide medical trainings. While some of these applications are used for medical purposes such as helping with diagnosis, preventing illnesses,

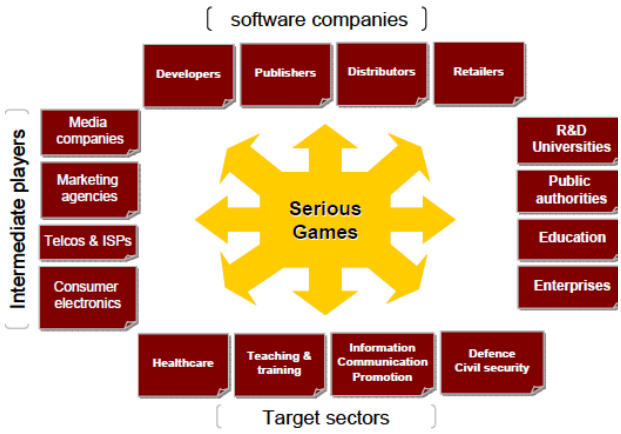


Fig 1. Categories of players involved in Serious Gaming

others offer training, fitness, physiotherapy and also advertising. To discuss the role of serious games in teaching and training sector, these games carried a large spectrum of aims such as teaching social science and physical science, management skills, foreign languages, team building skills etc. Thus, the games are not only presented to general public but also enterprises such as training specialist, manufacturers, military etc. Thirdly, public information and business communication sector is taken into consideration and it remarked that these games are used as an instrument to inform and communicate. When sector-specific serious games are analyzed, the purposes of past applications were diversified such as learning about renewable energy sources, promoting real estates, increasing awareness of flood protection and prevention policies etc. Finally, serious games that are devoted to defense and civil security sector are analyzed. It is found out that mainly, three features of serious games are used in military games: serving as recruitment tool, showcasing the armed forces and serving as training tool for military staff.

B. *Sample Games in Energy and Sustainability Field*

Bertoldi et al., asserted that intelligent education of young people which focuses on increasing awareness of the role of energy in societal level and household level and influences their energy related attitudes play an important role in terms of stimulating household energy conservation (as cited in Knol and Vries, 2000). Therefore, there has been many projects devoted energy efficiency, sustainability and renewable energy across the Europe and the serious gaming is considered as an innovative educational concept to reach young people. Compared to traditional educational tools, Bennet et al. asserted that serious gaming has higher potential because today's young people prefer visual information to textual, like to play games, are adapted and oriented to multiple media channels and highly active on social network sites (as cited in Knol and Vries, 2008). According to study of Grossberg et al., 53 games that all have the purpose of influencing behavior around energy efficiency and sustainability have been analyzed and it is found out that these serious games covered a wide range of solutions such as recycling, alternative energy, energy saving, utility efficiency etc. It is found out that these games are providing such common features like clear goals and rules, a compelling study, quick feedback and achievable short-term challenging tasks (as cited in Grossberg et al., 2015). In the study, 'Cool Choices' which is a is a nonprofit sustainability-focused serious game that is used in workplaces and schools and 'EnerCities' game which aims to reach out young people, especially secondary school students to alert and increase their awareness around energy saving and renewable energy are analyzed.

C. *Serious Games for Changing Behavior*

All in all, there are such actions required to take by individuals for energy efficiency and sustainability aspects and these actions are highly related to characteristics of individuals such as age, gender, preferences and attitudes. Therefore, serious games are recommended to change these behavioral patterns of people. It's stated that these games are used to support players in decision making, learning and aim to change behaviors whether the game is played by a single user or multiple users (Wood et al., 2014). To influence the occurrence or frequency of certain set of behaviors, empirical techniques are designed under applied behavior analysis (ABA) which is also called behavior modification (Due et al., 2014). According to researchers, behavior modification has been proven to be effective in different set of fields ranging from health, education to energy. However, these behavior modification methods are widely criticized in terms of resulting in unwanted side effects such as growing emotional disorders due to the punishments, arising hatred because of feeling of being manipulated through positive reinforcement and so on. Furthermore, these procedures of implementing behavior modification have been criticized by some researchers because a specific level of training is required and these trainings are time consuming and the procedures are not practical in terms of aligning with real life. In contrast to behavior modification which is made through imposing or removing a certain stimulus to lead behavioral change, serious games are claimed to achieve same results by creating an engaging experience using the elements of game Moreover, serious games relate such habit changes to positive emotional feedback and encourage positive behavior change in a pleasant and playful way (Schuller et al., 2013)

III. Methodolgy

Serious games can be of any genre, use any game technologies and be developed for any platforms (Kankaanranta et al., 2009). Serious games have common points with other games in term of owning its rules, accepting input from players, simulating behaviors and providing feedback within the context of the rules and behaviors (Michael et al., 2006). Therefore, design and development issues of serious games and entertainment games have similarities. Nevertheless, the primary design criteria in serious games highly differs since it is not "fun" but a serious purpose such as training, teaching etc. Thus, these games are more concerned with whether

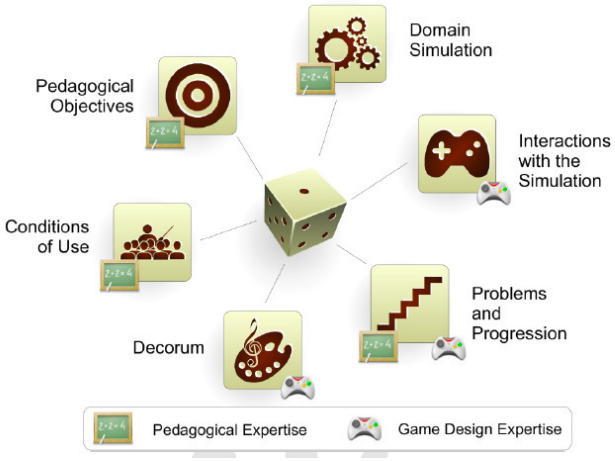


Fig 2. Six Facets of Conceptual Framework

Fig 3. The Game Board

or what the game is teaching and how the players can actually apply these learnt objects in the real life. It is discussed that serious games may be more entertaining but less capable in terms of knowledge acquisition if these were only designed by game developers. On the other hand, teacher or trainers might design such games that are powerful through educational purposes but these games would be lacking in motivating and engaging players (Marne et al., 2012). Thus, it is claimed that the conceptual framework which is called as 'Design Pattern Library' is created to facilitate the cooperation between various stakeholders in design stages of serious games. The stakeholders are grouped into two categories: the pedagogical experts and game experts. While pedagogical experts include educators, knowledge engineers, trainers and domain specialists; game experts consist of game developers, level designers, game producers, graphic and sound designers and so on (Marne et al., 2012). Therefore, six facets of the conceptual framework are used in the study and each facet is described by its title, a serious game design problem and a general solution and its experts. According to Marne et al.'s study, the six facets of the conceptual framework are given as: Pedagogical objectives, domain simulation, interactions with the simulation, problems and progression, decorum and conditions of use (2012). Indeed, the model helps to assign right experts to each of the six design areas of the serious games.

Briefly, pedagogical objectives aim to define the pedagogical content of the serious game and its general solution is to specify the knowledge model of the domain and educational purposes. The second facet, domain simulation, handles the problem of how to respond consistently the correct and incorrect actions taken by game players within a specific certain context. Thirdly, interactions with the simulation defines how to the engage the players when they are allowed to interact with the simulator. The forth facet, which is problems and progression, deals with which problems to give the players to solve and in which order these problems should be arranged. Decorum defines the types of multimedia or fun components of the game, that are not related to the domain simulation, to foster the player motivation. Finally, condition of use aims to define where, when, how and with whom the game is played.

IV. The Game

A. Game Description

The 'Enopoly' game aims to increase energy saving knowledge of students, to inspire them in terms of initiating energy-related projects in the campus and to increase their desire to continue a career path in energy sector. The game is developed as a board game and it takes place in the main campus of Istanbul Technical University as part of its scenario. As part of the scenario, players are granted with €1250 by the school rectorate and they aim to invest energy saving projects most wisely. The player who has the highest value of invested projects at the end of the game is selected as winner. The energy saving initiatives and projects at the campus are: Water purification system at the lake, energy conservation planting, bicycle stations, recycling center, campus farming, solar panels and wind turbines. Also, there are traditional buildings to invest such as dorm buildings, ITU Teknokent and Culture and Arts Union building.

Throughout the game, students are picking game cards when they land to specific areas and they lose or earn money based on scenarios related to their being investor of these projects. These areas are specified as below:

Green Reward: Player landing on Green Reward space collects €50 if she/he has investments in blue, green or yellow areas.

Think Tank: Players landing on Think Tank space lose a turn. During the turn, she/he must generate an energy-saving idea for the campus and share it with other places. Project ideas are pinned to the campus map which is on the game board.

Energy Fact: Players landing on the Energy Fact space must pick up a 'Energy Fact Card' and read it out.

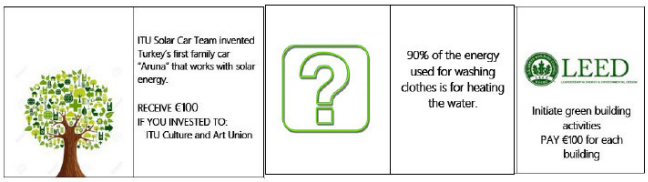


Fig 4. Sample Game Cards (Energy Fact, Environment, Investor and LEED cards in order)

Environment: Players landing on Environment space must pick up an 'Environment Card' and read it out. The player then either collects money from the bank or pays a fine.

Also, players landing on certain areas must pay a charge to the investor that is specified on investor card. Moreover, players earn more if they invest to same colored projects or if they have LEED certification investment for traditional buildings.

B. Game Flow

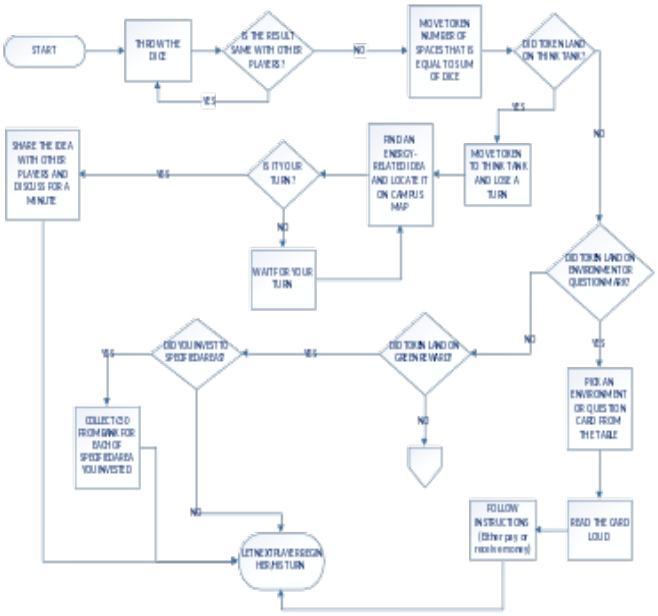


Fig 5. First Part of Game Flow

V. Case Study

The game was first played with 10 different students in a classroom and the player groups were composed of 3 to 4 students. Before playing the game, a survey was handled to the students. Therefore, students responded to survey questions without knowing the concept of the game. The reason why students were not introduced to the game was to minimize any bias effects. With these survey questions, it's aimed to measure students' current knowledge, concerns, motivations about energy-related topics, projects and jobs. As certain knowledge, concerns and intrinsic motivations lead behavioral patterns of individuals, the survey results will be used to measure players' current situation. After the collecting pre-game survey answers, group of students were

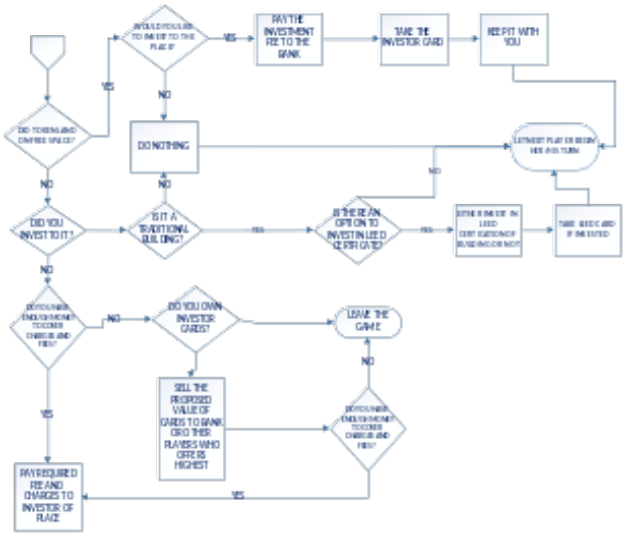


Fig 6. Second Part of Game Flow

introduced to the Enopoly game. At the end of the game, after-game survey questions were handled to players. With these survey questions, it's aimed to measure game's effect on players about their knowledge, concerns, motivations about energy-related topics. Therefore, a comparison between survey results before and after playing the game will be used as an indicator to analyze effect of the game on players

Table 1
Comparison of Survey Results

	Pre-Game Results		**After-Game Results**	
	Mean	Std. Dev.	Mean	Std. Dev.
Interest in learning about energy saving and sustainability	4.3	1.05	4.4	0.51
Awareness about lowering energy usage as an individual	4	0.66	4.5	0.52
Awareness about responsibility to initiate energy saving projects at campus	3.6	0.96	4.4	0.69
Concerns about environment	4.7	0.48	4.4	0.51
Consideration of a career path in energy sector	2.8	1.03	3.4	0.96
Desire to get a job in energy sector	3.2	0.63	3.6	0.84
Desire to initiate energy saving projects	3.6	0.84	4.1	0.73

VI. Results and Discussion

By taking pre-game and after-game survey results into consideration, a comparison is made to measure the effect of Enopoly game on players. According to the results, it's seen that players' interest in learning about energy saving and sustainability increased slightly and awareness about lowering energy usage as individual increased considerably from 4 to 4.5. As the game aims to stimulate the awareness and attitude of university students relating to energy saving and sustainability, the results were positively aligned with it. Also, players' awareness about their responsibility to initiate energy saving projects at campus has risen dramatically from 3.6 to 4.4. Throughout the game play, it was observed that players highly enjoyed Think-Tank space as they had opportunity to voice campus energy-saving ideas with other players and have such discussions.

Players indicated that they are already concerned about environment at the before the game play. Therefore, there has not been significant change related to concerns. Additionally, players proved a higher motivation and desire to follow a career path in energy sector after playing the game. Also, the game stimulated players' desires to initiate energy saving projects; there has been a significant increase from 3.6 to 4.1 at the end of the game. It is found that in contrast to behavior modification which is made through imposing or removing a certain stimulus to lead behavioral change, serious games are powerful to achieve same results by creating an engaging experience using the elements of game. Also, it is claimed that serious games relate such habitual changes to positive emotional feedback and encourage positive behavior change in a pleasant and playful way (Schuller et al., 2013) According to the players' feedbacks and observations, Enopoly game made such engaging experience to the players with energy saving and sustainability topics in a playful way. Also, data analyses and student reflections shown that Enopoly has such a power to effect players' knowledge, motivation and intrinsic desires on energy- related topics, projects and jobs.

References

[1] Bensch, I. (2012). Miron Construction's iChoose Game. *Energy Center of Wisconsin*.

[2] Consumption of energy. (n.d.). Retrieved January 15, 2016, from http://ec.europa.eu/eurostat/statistics-explained/. Eurostat.

[3] Due, J., Feng, Y., & Zhou, C. (2014). *Gamification for Behavior Change of Occupants in Campus Buildings* (Unpublished master dissertation). Duke University.

[4] Energy Statistics. (n.d.). Retrieved March 15, 2016, from http://www.tuik.gov.tr/PreTablo.do?alt_id=1029 Türkiye İstatistik Kurumu (TUİK).

[5] Grossberg, F., Wolfson, M., Mazur- Stommen, S., Farley, K., & Nadel, S. (2015). Gamified Energy Efficiency Programs. *American Council for an Energy-Efficient Economy*.

[6] How is electricity used in U.S. homes? (n.d.). Retrieved February 11, 2016, from https://www.eia.gov/tools/faqs/faq.cfm?id=96

[7] U.S. Energy Information Administration - EIA - Independent Statistics and Analysis

[8] How We Use Energy. (n.d.). Retrieved April 3, 2016, from http://needtoknow.nas.edu/energy/energy-use/. The National Academies of Sciences

[9] Kankaanranta, M., & Neittaanmäki, P. (2009). *Design and use of serious games*. Dordrecht: Springer.

[10] Klemetti, M., Taimisto, O., & Karppinen, P. (n.d.). The Attitudes of Finnish School Teachers Towards Commercial Educational Games. *Design and Use of Serious Games,* 97-105.

[11] Knol, E., & Vries, P. W. (2011). EnerCities - A Serious Game to Stimulate Sustainability and Energy Conservation: Preliminary Results. *ELearning Papers,* 25th ser. Retrieved from http://ssrn.com/abstract=1866206

[12] Laamarti, F., Eid, M., & El Saddik, A. (2014). An overview of serious games. *International Journal of Computer Games Technology*. Hindawi Publishing Corporation.

[13] Lebram, M., Backlund, P., Engström, H., & Johannesson, M. (2009). Design and Architecture of Sidh – a Cave Based Firefighter Training Game. *Design and Use of Serious Games,* 19-31.

[14] Marne, B., Wisdom, J., Huynh-Kim-Bang, B., & Labat, J. (2012). The Six Facets of Serious Game Design: A Methodology Enhanced by Our Design Pattern Library. *21st Century Learning for 21st Century Skills, 7563*, 208-221.

[15] Michael, D., & Chen, S. (2006). *Serious Games: Games That Educate, Train and Inform*. Boston, MA: Thomson Course Technology.

[16] Michaud, L., Alvarez, J., Alvarez, V., & Djaouti, D. (2010). *Serious Games: Training and teaching, Healthcare, Security and defence, Information and communication* (2nd ed.). Montpellier: IDATE.

[17] National Energy and Education Development. (2016). *Saving Energy at Home and School* [Brochure]. Manassas, VA: Author. Student and Family Guide

[18] Schuller, B. W., Dunwell, I., Weninger, F., & Paletta, L. (2013). Serious Gaming for Behavior Change: The State of Play. *IEEE Pervasive Comput. IEEE Pervasive Computing, 12*(3), 48-55. doi:10.1109/mprv.2013.54

[19] OECD. Household Behaviour and the Environment: Reviewing the Evidence. (2008). Paris.

[20] Jonathan Kaplan. (2010). *Eat Green: Our everyday food choices affect global warming and environment* [Brochure]. Author.

[21] Water Program. (n.d.). Retrieved April 10, 2016, from http://www.gracelinks.org/824/water-program

[22] Wood, G., Horst, D. V., Day, R., Bakaoukas, A. G., Petridis, P., Liu, S., Pisithpunth, C. (2014). Serious games for energy social science research. *Technology Analysis & Strategic Management, 26*(10), 1212-1227. doi:10.1080/09537325.2014.978277

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 233–238
DOI: 10.15439/2016F584
ACSIS, Vol. 9. ISSN 2300-5963

Big Data solutions in cloud environment

Maciej Pondel
Wrocław University of Economics
Komandorska 118/120
53-345 Wrocław, Poland
Email: maciej.pondel@ue.wroc.pl

Jolanta Pondel
University of Business in Wroclaw
Ostrowskiego 22
53-238 Wroclaw, Poland
Email: Jolanta.pondel@handlowa.eu

***Abstract*—Current business faces new challenges that require modern and adjusted IT models. Authors of this paper try to identify and indicate selected challenges that are addressed by by cloud computing concept and Big Data solutions. Authors of this paper concentrate od Microsoft Azure cloud offering mainly in area of Big Data and they want to prove that development of Big Data solutions in cloud environment can be efficient from financial and functional perspective.**

I. Introduction

Nowadays organizations are facing new challenges due to emerging business models. They operate on a market where competitors provide customers with more sophisticated and advanced services. Enterprises need to benefit from modern technologies if they want to reach or even overtake competitors. There is number of examples of companies that invented new service or product and seriously changed the way the market works. Examples of Uber, Netflix, Spotify, Airbnb or WhatsApp show that basing on communication technologies and efficient data processing and usage one can build business worth much more than the one working in traditional model. Of course market leaders are forced to continuously streamline their business basing on 2 main pillars: delivery of unique value to their clients and customers and controlling efficiency of business processes – mainly based on cost-cutting. Companies like Microsoft, Google, Amazon, Facebook are in a process of continuous improvement of business strategies, models and operational processes to keep their current position or expand the business.

II. Business and IT Challenges

There are several IT challenges that companies need to consider in current business. Among them we can distinguish:

- Innovation agility [1],[2] – understood as flexibility of IT architectures and business approaches to quickly deliver value, evaluate efficiency and market potential and if necessary scale up the innovation or withdraw with reasonable incurred expenses
- Interoperability and microservices architectures [3], [4] – which delivers IT architectures that are from IT perspective understandable, maintainable, scalable. Loosely coupled components are reusable. They collectively provide the complete functionality of a large software application. The services cooperate by exchanging data and information with other services without any human interaction. The services should be black boxes with precisely defined input parameters and output results. From business perspective such approach allows to efficiently create innovative services for clients that are composition of carefully selected microservices delivered by various vendors that we consider to be most efficient, proficient and capable. Various software applications are capable to cooperate because of interoperability between different programming languages, systems that allow integration of services. Varous types of resources are federated, what allows transparently mapping multiple autonomous resources to be treated by users as one federated resource.
- Mobile interfaces [5],[6] – modern requirements regard ability to complete interaction with software application using all possible devices. Users or customers spend more and more time in Internet using not only smartphones or tablets but also wearable devices providing discrete interfaces that allow control of mobile devices through subtle gestures in order to gain social acceptance. If company provides customer/user with the capabilities empowering use various possible devices and ways of communication with their service, it will be considered to be more

attractive but also impressive and efficient than others.

- Consumerization [6] of business activities expressed mainly in BYOD approach. It addresses the adoption of consumer devices and applications in the workforce. Employees bring computer tablets and smartphones into the workplace and harness social media applications and special purpose apps for their work lives. Such behavior if not properly managed may impact enterprise security but also can increase productivity of individuals and teams. The effects of consumerization are considered to be a major driver that redefines the relationship between employees (in terms of consumers of enterprise IT) and the IT organization [7]. Examples of consumerization we can observe in a phenomenon of existence iPhone and iPad in business scenarios, the usage of Facebook in companies and popularity of such application like Yammer or Hipchat that are based on Facebook. Another example are Google-like enterprise search indexing internal business content and documents in company that seem to be efficient in knowledge management scenarios.
- Internet of things measurability. The advancements and convergence of micro-electro-mechanical systems (MEMS) technology, wireless communications, and digital electronics has resulted in the development of miniature devices having the ability to sense, compute, and communicate wirelessly in short distances. These miniature devices called nodes interconnect to form a wireless sensor networks (WSN) and find wide ranging applications in environmental monitoring, infrastructure monitoring, traffic monitoring, retail, etc. [8]. IoT approach is not narrowed only to connecting things such as devices, machines etc. to the Internet, but it also allows them to interact with each other, exchange specific data, and complete some tasks without human interaction. Machine-to-machine (M2M) communication is nothing new, but this way of using sensors and wireless connection is revolutionary. Basic example of M2M operation is when sensors gather data, send it to a network either wirelessly or via cable connection, where its directed to a central server. We are able to create a wide range of IoT based business scenarios that give a real benefits to end users. We also have to be aware that all those signals and messages exchanged by IoT devices occur to be a valuable data that can be gathered and used, for example for decision making or enhancing operations. The amount of data generated by IoT devices is grooving rapidly. The figure 1 presents forecasts regarding number of connected devices in use globally. Assuming that part of them produce data that is worth storing and analyzing we can conclude that it is impossible to build efficient analytical system basing on relational databases and data warehousing technologies. Due to anticipated efficiency issues dig data technologies have to be involved in the processing of Internet of Things originated data.

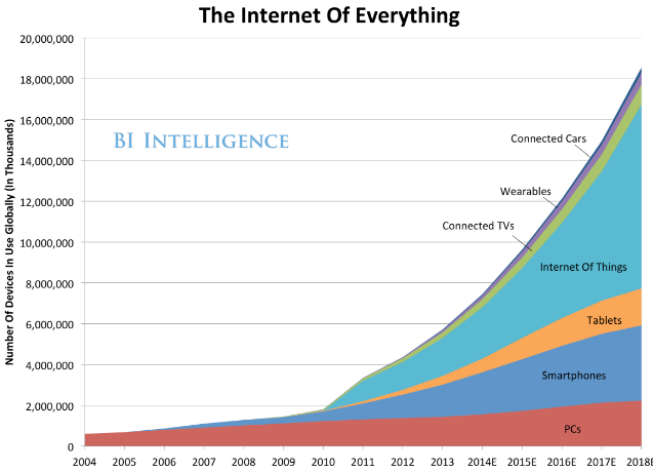


Figure 1 Forecasts for the entire Internet-connected ecosystem. Source: BI Intelligence [9]

Mentioned challenges are addressed mainly by two emerging IT concepts that are significant nowadays and affect business models. They are Big Data and Cloud Computing. Formally separable concepts in some areas are strongly related and their association can deliver real profits for users. Authors are aiming to present both concepts basing on an example of Microsoft Azure offering that is one of the most advanced platforms addressing mentioned challenges.

II. CLOUD COMPUTING OFFERING

Of course Microsoft platform is not the only existing. There are competitive platform provided by Amazon (Amazon Web Services), Google Cloud Platform provided by Google. Authors decided to mention those 3 because according to Gartner's report called Magic Quadrant they are worldwide leaders. Magic quadrants evaluates solution in 2 dimensions [10]:

- Ability to Execute - Gartner analysts evaluate technology vendors on the quality and efficacy of the processes, systems, methods or procedures that enable IT providers' performance to be competitive, efficient and effective, and to positively affect revenue, retention and reputation. The following criteria were evaluated: Product/Service, Overall Viability, Sales Execution/Pricing, Market

Responsiveness/Record, Marketing Execution, Customer Experience, Operations.

- Completeness of Vision - Gartner analysts evaluate technology vendors on their ability to articulate logical statements convincingly about current and future market direction, innovation, customer needs and competitive forces, as well as how they map to Gartner's position. The following criteria were taken into consideration: Market Understanding, Marketing Strategy, Sales Strategy, Offering (Product) Strategy, Business Model, Vertical/Industry Strategy, Innovation, Geographic Strategy.

Gartner prepared separate quadrants in the following categories:

- Infrastructure as a Service (IaaS) is a type of cloud computing service; it parallels the infrastructure and data center initiatives of IT. Cloud compute IaaS constitutes the largest segment of this market (the broader IaaS market also includes cloud storage and cloud printing)[10]. Figure 2 presents IaaS Quadrant.
- Platform as a Service (PaaS) is a cloud computing model that delivers applications over the Internet. In a PaaS model, a cloud provider delivers hardware and software tools -- usually those needed for application development -- to its users as a service. A PaaS provider hosts the hardware and software on its own infrastructure.[12]

Figure 2.Magic Quadrant for Cloud Infrastructure as a Service, Worldwide
Source [10]

Figure 3. Magic Quadrant for Cloud Platform as a Service, Worldwide
source [11]

Regarding IaaS offering Amazon Web Services is considered to be worldwide leader. Microsoft Azure is located on the 2nd place and Google on 3rd. In terms of PaaS offering Salesforce is acknowledged to be a leader. PaaS model is only delivered by Salesforce that is why the company was not mentioned by authors as cloud provider. Microsoft and Google have 2nd and 3rd place.

The aim of authors is not to evaluate the proficiency of every cloud provider. Microsoft Azure was chosen because of its existence in the cloud providers rankings and because of other research works conducted by authors in that platform.

The main advantages of using cloud services boil down to:

- Savings in terms of initial expenses on infrastructure and platform creation. Cloud services require payments for the resources that company really utilizes. In most cases there are no upfront costs assigned to hardware and software licenses purchase.
- Scalability - If application needs a large computing power for relatively short time the cloud can provide resources dynamically what is financially efficient. It is also available to allocate a large amount of resources in a relatively short time what is impossible in regards to on-premises environment.
- The operational costs are reduced because they are spread over a number of clients. In most cases it is more efficient to subscribe cloud services then to provide independently such constituents like energy, air conditioning, renting premises, security, maintenance, networking, management and many others.
- Wide range of well standardized services that can enrich capabilities of planned solution. In MS Azure the number of new services increased by 500 during

one year. Of course some of them are available in a preview edition (what means they have not been fully stable yet) but majority of them is available to use and they are relatively easy to integrate with other solutions.

There are also some boundaries related with cloud services that company needs to accept before starting usage of cloud services:

- Standardization – public cloud providers have to standardize their services in order to be able to maintain their quality. If clients defines individual requirements because of some special resources demands or he needs individual services or contracts – cloud vendor most probably will not be able to fulfill client's requirements.
- Limited SLA – in most cases cloud providers offer reasonable SLA (MS Azure offers from 99,9% to 99,95% SLA depending on service and chosen architecture). If client requires higher value – in most cases it is impossible to assure such quality.
- Distance between data and their user. If we store data in Local Area Network we have a high speed access. If data are stored in cloud we reach them through Internet connection that in most cases is slower than LAN. If we have a problem with Internet reliability we can be isolated from our data or applications.

Authors do not mention the security as a boundary of cloud offering because all leading vendors addressed the issue and are able to assure security on much higher level than it is possible in on premises infrastructure.

Indicated advantages in most cases address the challenges mentioned earlier. Cloud eliminates entrance barriers that supports innovation agility. A number of available, well standardized services support the benefits of interoperability and microservices architectures.
Scalability is curtail in every Internet of Things solution. We cannot predict precisely the final number of connected devices and amount of produced data that is why such solutions require a flexible resource allocation what is beneficial from performance perspective.

III. BIG DATA ECOSYSTEM OVERVIEW

As it was mentioned current business generate and possess a huge amount of data created during their operational activity. The data is stored in various IT systems in miscellaneous formats and very often in different locations. Continuous development of Computer Sciences and Information Technologies is increasing the level of knowledge concerning methods of designing and creating databases and processing the data that give us new possibilities and opportunities in acquiring information and knowledge increasing the business performance.

Big data concept is not aimed to deliver a business critical transactional systems. It is more about performing analytics.

Recently big data and big data analytics have been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data [13] . The challenge of big data is to manage a large volume of data with optimal processing time [14].

Describing Big Data architecture we have to start from MapReduce concept is one of the most important techniques, and currently the preferred choice of cloud providers, for providing cloud-based data analysis services [15]. This platform invented by Google allows distribute processing on large number of computers. It is basing on 2 steps. Mapping is a step splitting the complex task into subtasks and reduce is about aggregation of results to combine them into one answer. Map function processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function merges all intermediate values associated with the same intermediate key [16],[17].

Basing on a MapReduce concept Apache™ Hadoop® platform emerged. The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules[18]:

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

Basing on Apache Hadoop interesting projects are developed like:

- Cassandra™: A scalable multi-master database with no single points of failure.
- HBase™: A scalable, distributed database that supports structured data storage for large tables.
- Hive™: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- Mahout™: A Scalable machine learning and data mining library.

- Pig™: A high-level data-flow language and execution framework for parallel computation.
- Spark™: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.

IV. BIG DATA IN MICROSOFT AZURE

Microsoft is an active contributor to the Apache Software Foundation development effort. Their engineers works, committing code and driving innovation in partnership with the open source community across a range of Hadoop projects. Microsoft understand benefits of contributing Apache Hadoop development because they include most important Hadoop Components into MS Azure scope.

Most important Azure service called Azure HDInsight is a 100% Apache Hadoop-based service in the Azure cloud. It offers all the advantages of Hadoop, plus the ability to integrate with Excel, your on-premises Hadoop clusters and the Microsoft ecosystem of business software and services [19]. Azure HDInsight includes the most important modules and components like:

- MapReduce
- Pig
- Hive
- HBase
- Storm
- Spark
- RServer.

Azure also includes 3rd party solutions basing on Hadoop concept. They are:

- Cloudera Enterprise Data Hub
- Hortonworks Sandbox on Azure

Why to implement Hadoop in the cloud? Deploying Hadoop on-premises requires hardware and skilled Hadoop experts to set up, tune and maintain them. Cloud service possess preconfigured platforms that we can launch in minutes without up-front costs. Most big data tasks require individual tasks of data processing and after it is finished we can shut down the platform and stop incurring expenses.

If we want to initiate Hadoop cluster in Microsoft Azure platform we have to specify:

1. Type o a cluster. We have to choose from: Hadoop, Storm, HBase, Spark and R Server (2 latest existing in a preview edition). We also need to define Operationg system (Linux or Windows) and verision of chosen software. Confoguration of a cluster type is presented on figure 4.

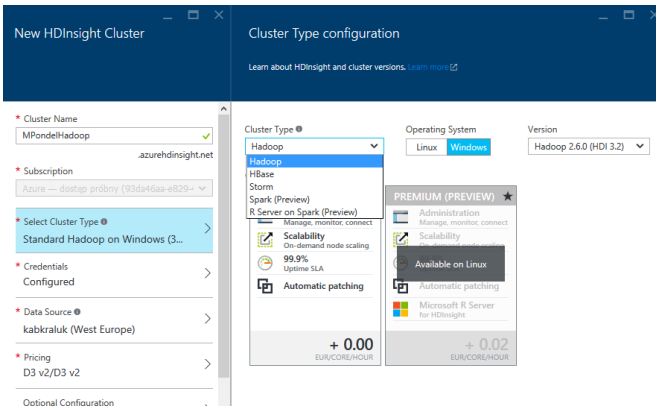


Figure 4. Configuration of cluster type
Source: own work on Azure Portal

2. In a next step we have to provide credential – admin password and an name and password of a remote desktop user
3. We have to indicate the storage account for the cluster
4. We have to choose the pricing tier. We specify the number of nodes (computers in cluster) and their parameters. This step is presented on a figure 5.
5. We can also define Optional configuration regarding:
    a. Virtual network
    b. External Metastores
    c. Script Actions
    d. Linked Storage Accounts
6. We should also allocate our cluster to resource group what determines the localization of servers.

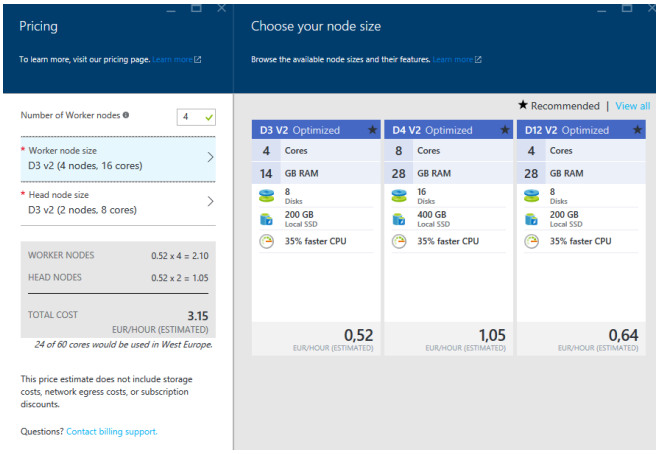


Figure 5. Definition of a pricing tier.
Source: own work on Azure Portal

After initiation of a process crating platform we had to wait about 20 minutes for Hadoop cluster to be created. We can connect using remote desktop and perform operations.

CONCLUSIONS

New business challenges require modern and adjusted IT models. Some of challenges are addressed by cloud computing concept. In modern business we need analytical tools that are able to perform analytical tasks efficiently. Big

Data solutions can support those business needs. Authors of this paper wanted to prove that building Big Data solutions in cloud environment in some cases can be more efficient and give additional value supporting enterprise business.

REFERENCES

[[1] Boyer, M. J., & Mili, H. (2011). Agile business rule development (pp. 49-71). Springer Berlin Heidelberg
[2] Wilson, K., & Doz, Y. L. (2011). Agile innovation: A footprint balancing distance and immersion. California Management Review, 53(2), 6-26.
[3] Maciaszek, L. A. (2008). Building Quality into Web Information Systems. In WEBIST (1).
[4] Pondel, M. (2013, September). Business Intelligence as a service in a cloud environment. In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on (pp. 1281-1283). IEEE.
[5] Rico, J., & Brewster, S. (2010, April). Usable gestures for mobile interfaces: evaluating social acceptability. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 887-896). ACM.
[6] Harris, J., Ives, B., & Junglas, I. (2012). IT Consumerization: When Gadgets Turn Into Enterprise IT Tools. MIS Quarterly Executive, 11(3).
[7] Niehaves, B., Köffer, S., & Ortbach, K. (2012). IT consumerization–a theory and practice review
[8] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645-1660.
[9] BI Intelligence, Business Insider Research http://www.businessinsider.com/the-internet-of-everything-2014-slide-deck-sai-2014-2
[10] Leong, L., Toombs, D., & Gill, B. (2015). Magic Quadrant for Cloud Infrastructure as a Service, Worldwide. Analyst (s), 501, G00265139.
[11] https://azure.microsoft.com/pl-pl/blog/microsoft-the-only-vendor-named-a-leader-in-gartner-magic-quadrants-for-iaas-application-paas-cloud-storage-and-hybrid/
[12] http://searchcloudcomputing.techtarget.com/definition/Platform-as-a-Service-PaaS
[13] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.
[14] Business Opportunity Detection in the Big Data
[15] Ranjan, R., Georgakopoulos, D., & Wang, L. (2016). A note on software tools and technologies for delivering smart media-optimized big data applications in the cloud. Computing, 98(1-2), 1-5.
[16] Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.".
[17] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
[18] http://hadoop.apache.org/
[19] https://azure.microsoft.com/en-gb/solutions/hadoop/

22nd Conference on Knowledge Acquisition and Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference “Knowledge Acquisition and Management". The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we’ve joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The aim of this event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

TOPICS

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

EVENT CHAIRS

- **Hauke, Krzysztof,** Wroclaw University of Economics, Poland
- **Nycz, Malgorzata,** Wroclaw University of Economics, Poland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland
- **Pondel, Maciej,** Wroclaw University of Economics, Poland

PROGRAM COMMITTEE

- **Andres, Frederic,** National Institute of Informatics, Tokyo, Japan
- **Chmielarz, Witold,** Warsaw University, Poland
- **Christozov, Dimitar,** American University in Bulgaria, Bulgaria
- **Helfert, Markus,** Dublin City University, Ireland
- **Jan, Vanthienen,** Katholike Universiteit Leuven, Belgium
- **Jelonek, Dorota,** Faculty og Management of Czestochowa University of Technology
- **Kania, Krzysztof,** University of Economics in Katowice, Poland
- **Kayakutlu, Gulgun,** Istanbul Technical University, Turkey
- **Korczak, Jerzy,** Wrocław University of Economics, Poland
- **Ligęza, Antoni,** AGH University of Science and Technology, Poland
- **Mach-Król, Maria,** University of Economics in Katowice, Poland
- **Mercier-Laurent, Eunika,** University Jean Moulin Lyon3, France
- **Nalepa, Grzegorz J.,** AGH University of Science and Technology, Poland
- **Sobińska, Małgorzata,** Wroclaw University of Economics, Poland

- **Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Vasiliev, Julian,** University of Economics in Varna, Bulgaria
- **Zhelezko, Boris,** Belorussian State Economic University, Belarus
- **Zhu, Yungang,** College of Computer Science and Technology, Jilin University, China

ORGANIZING COMMITTEE

- **Marciniak, Katarzyna,** Wroclaw University of Economics, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 241–244
DOI: 10.15439/2016F561
ACSIS, Vol. 9. ISSN 2300-5963

Balance recognition on the basis of EEG measurement

Natalia Tusk
Gdańsk University of Technology
ul. Narutowicza 11/12
80-233 Gdańsk, Poland
email: tusk.nat@gmail.com

Artur Poliński
Gdańsk University of Technology
ul. Narutowicza 11/12
80-233 Gdańsk, Poland
email: apoli@biomed.eti.pg.gda.pl

Tomasz Kocejko
Gdańsk University of Technology
ul. Narutowicza 11/12
80-233 Gdańsk, Poland
email: tomkocej@pg.gda.pl

***Abstract* — Although electroencephalography (EEG) is not typically used for verifying the sense of balance, it can be used for analysing cortical signals responsible for this phenomenon. Simple balance tasks can be proposed as a good indicator of whether the sense of balance is acting more or less actively. This article presents preliminary results for the potential of using EEG to balance sensing. The results are not unequivocal and further research is required.**

I. Introduction

SENSE of balance is an extremely important ability that allows humans to maintain an upright posture or to remain stable in many advanced motion processes. However, the organs responsible for the sense of balance partially lose their function with age, which is a major problem in the aging population. This problem also concerns younger people with dysfunctions or diseases of organs responsible for equilibrioception [1].

People over 65 years old lose balance significantly more frequently than younger people. In this group, one in every three persons experiences one or more falls due to balance loss each year [1]. This can lead to serious injuries and eventually become a life threat. Even if the consequences are not too severe, living with a poor sense of balance can be a serious inconvenience, causing a fear of falling. Other long-term consequences include weak self-reliance, low social and behavioural activity and more limited mobility. For this reason, there is a raising need for sensor-based multimodal systems that provide care and support for older people [2].

Taking into consideration the fact that nowadays the population is aging, sense of balance examinations may help to diagnose and in consequence treat balance disorders. This could raise the standard of living of older or ill people, and prevent many dangerous accidents from happening.

Physiologically, there are a few organs responsible for balance control. The signals' source for the sense of balance originates from the vestibular system, the eyes and proprioceptors. These signals are analysed by the central nervous system, mainly the cerebellum and the brain stem. The involvement of the cerebral cortex in balance control has also been proven [3], [4]. For this reason, examining the human brain and its functions is of great importance in relation to balance.

Most basic examinations of human balance are neurological tests such as the finger-to-nose test, the Romberg test, the Unterberger and Fukuda stepping test, the Hautant's test or gait analysis [5]. Performing diagnostic procedures with computer assistance helps to gather and interpret more valuable data such as posturography, craniocorpography or the Kinect supported analysis of physical tests. Taking into account the fact that balance loss causes involuntary eye movements called nystagmus, diagnostic procedures can also include: electronystagmography (ENG), electrooculography (EOG), videonystagmography (VNG), infrared reflection oculography (IROG), and the search coil technique. Because one of the most important parts of the equilibrium organ in the human body is the vestibular apparatus, the last type of methods used are audiometric tests, mainly electrocochleography (ECOG) and Brainstem Auditory Evoked Potentials. The least frequently used method for analysing the human balance is electroencephalography (EEG). This is because it is the most high-level method. The cerebral cortex is responsible for analysing all neural signals, including balance sensing. This is why it becomes more and more popular in terms of balance examination [3]. It requires advanced signal processing to obtain the diagnostic information from the EEG signal. The most common technique of EEG signal analysis is the Fourier transform. Such an approach allows analysing different frequency bands separately.

Calculating the Fourier transform of the signal enables the use of Power Spectral Density (PSD) which reflects the distribution of signal power over frequency [6]. PSD has been used for identifying cortical activity changes on the basis of EEG [3].

The aim of the study was to verify if using EEG was possible to differentiate the one and two feet standing position which required more body balancing than the stable both feet posture.

This work has been partially supported by statutory funds of the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology.

II. METHODS

A. Design of the test and examination conditions

The tests consisted of two EEG measurement series for each patient. The first one was conducted for a patient who was standing with both feet on the ground maintaining an upright standing posture. The second one was conducted for a patient standing on one foot and also trying to maintain an upright standing posture.

Each measurement series lasted approximately three minutes – all data were cut to the same length during the analysing process. All other external stimuli, apart from one or two feet standing position and vision were excluded.

B. Examined subjects

Seven volunteers were recruited for the tests, four men and three women. All of them were young - between 20 and 35 years of age - and their sense of balance was supposedly ordinary. They did not have any documented disorders that might have influenced it. Their anatomy also did not indicate any pathological disorders that might have resulted in postural deviations. In order to compare the results for open and closed eyes, some measurement series were conducted for open (3 subjects, 1 man and 2 women, patients number 1-3) and some for closed eyes (4 subjects, 3 men and 1 woman, patients number 4-7).

C. Measurement system structure

The BioSemi EEG system was used during the measurements [7]. It is a multi-channel, high resolution biopotential measurement system for research applications. The system consisted of: 16 active electrodes as well as two reference electrodes, headcap, AD-box, which is an analog-to-digital converter receiving and amplifying the signals coming from the electrodes, USB receiver, which converts the digitized signal into USB output, PC workstation directly connected to the USB receiver, and dedicated software for EEG data analysis.

The electrodes were located in accordance with the manufacturer's suggestions for 16 active electrodes' application. Their location was compatible with standardized 10-20 system for EEG measurement (Fig. 1).

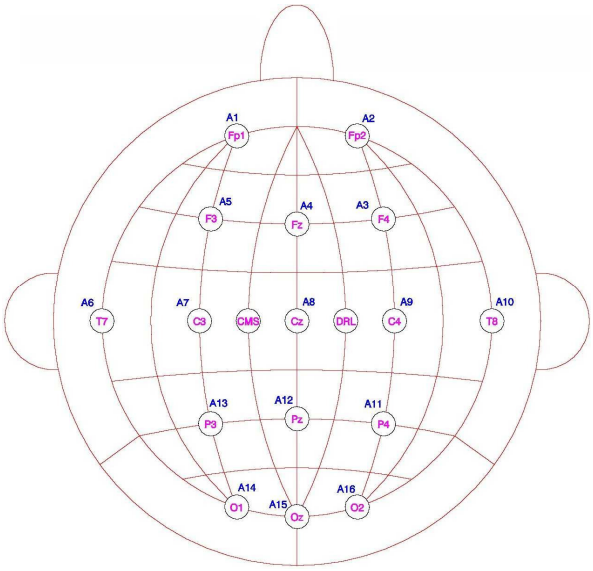


Fig. 1 Electrodes' location during the experiments [7].

The software used for EEG data analysis was ActiView. It is a LabVIEW-based program for the acquisition of EEG signals. The EEG sample rate was set to 2048 Hz. The software also allowed the electrodes' impedance to be checked. It was checked for every subject and more gel was applied in case it was too big.

D. Data analysis

All data were processed using the Matlab environment. First, the magnitude squared coherence (MSC) [8] for the two signals (one and two feet standing positions) was calculated using originally and down sampled data

$$MSC = \left|C_{xy}(\omega)\right|^2 = \frac{\left|G_{xy}(\omega)\right|^2}{G_{xx}(\omega)G_{yy}(\omega)} \tag{1}$$

where $G_{xx}(\omega)$, $G_{yy}(\omega)$ are the power spectra of these signals, and $G_{xy}(\omega)$ is the cross-power spectrum. These parameters can be obtained by means of the Fourier transform

$$G_{xx}(\omega) = \left|X(\omega)\right|^2 \tag{2}$$

$$G_{yy}(\omega) = \left|Y(\omega)\right|^2 \tag{3}$$

$$G_{xy}(\omega) = X(\omega)Y^*(\omega) \tag{4}$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of one and two feet standing positions' EEG signals. The magnitude squared coherence was calculated to find the most significant frequency band for further analysis. The band selection was also supported by results presented in [3]. Finally, the beta rhythm defined as 13 – 19 Hz and sigma defined as 30 – 40 Hz frequency range were chosen.

Next, the power for the selected bands of the signal was calculated using the Fourier transform. Since we are interested in the reaction to increased balance requirements, time changes of the power were estimated. The reference signal was taken as the average power for the selected bands calculated for the two feet standing position.

In the case of the one foot standing position, the power was calculated for 60 s and 1 s time windows shifted through the whole measurement time.

The measurements were conducted with open and closed eyes since in the second case the balance preservation is more complicated.

The power for the selected frequency band was calculated using PSD

$$PSD(X) = \left| X(\omega) \right|^2 \tag{5}$$

where $X(\omega)$ is the Fourier transform of the input $x(t)$ signal.

III. Results

In the following figures the relative power versus time is shown. These were normalized in relation to the mean value of both feet standing power for the selected channel.

The relative power for the beta band of channel number 10 for the one foot standing position of patient number 4 for 60 s and 1 s averaging windows is shown in Fig. 2. To better show the difference, the power was normalized to the power of 60 s window.

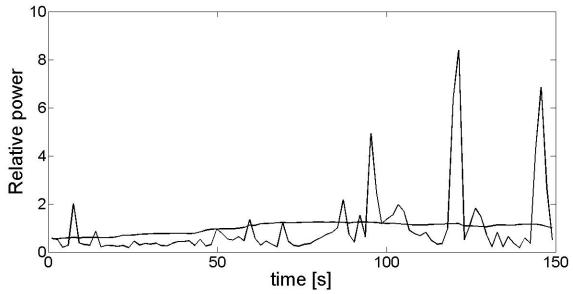


Fig. 2 Example of the length of averaging window influence.

The relative power for the beta band of channel number 10 for the one foot and two feet standing positions of patient number 3 for 60 s averaging window is shown in Fig. 3.

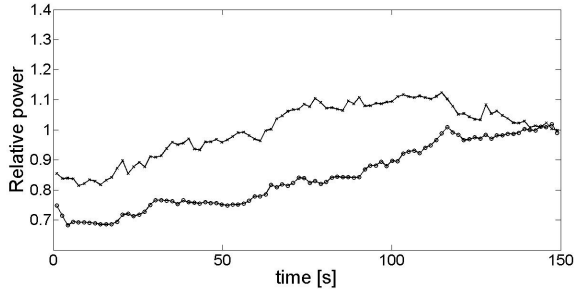


Fig. 3 Example of relative power for the one foot (x) and two feet (o) standing positions for open eyes.

The relative power for the beta band of channel number 10 for the one foot and two feet standing positions of patient number 5 for 60 s averaging window is shown in Fig. 4.

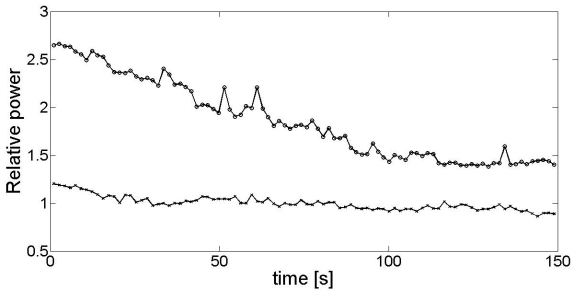


Fig. 4 Example of relative power for the one foot (x) and two feet (o) standing positions for closed eyes.

The relative power for the beta band of channel number 9 for the one foot and two feet standing positions of patient number 7 for 60 s averaging window is shown in Fig. 5.

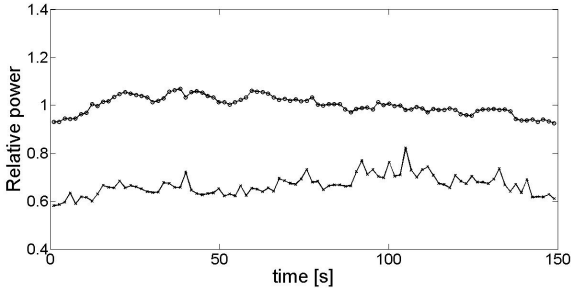


Fig. 5 Example of relative power for the one foot (x) and two feet (o) standing positions for closed eyes.

The relative power for the sigma band of channel number 9 for the one foot and two feet standing positions of patient number 7 for 60 s averaging window is shown in Fig. 6.

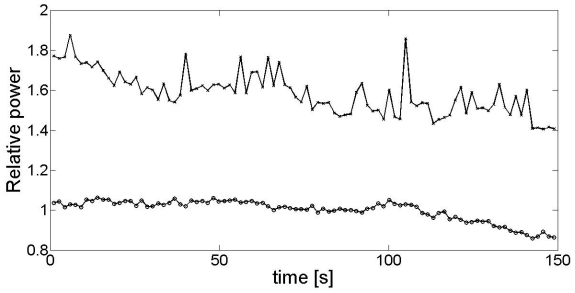


Fig. 6 Example of relative power for the one foot (x) and two feet (o) standing positions for closed eyes.

The relative power for the beta band of channel number 9 for the one foot and two feet standing positions of patient number 4 for 60 s averaging window is shown in Fig. 7.

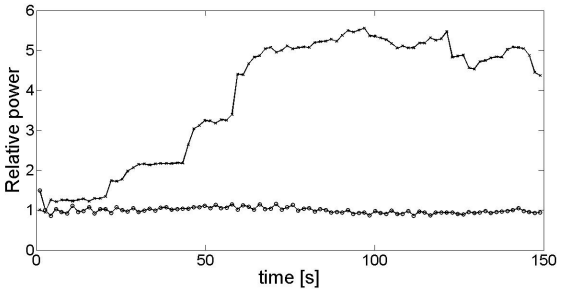


Fig. 7 Example of relative power for the one foot (x) and two feet (o) standing positions for closed eyes.

IV. Discussion

The measurement results are not fully consistent. The number of examined subjects might not have been sufficient. Increasing the number of examined subjects may lower the discrepancy between data.

The coherence did not clearly indicate the best frequency band for analysis - neither for any examined electrode, nor for any examined subject. This may be due to complications in its calculations. Because of this, further comparison was conducted for beta and sigma frequency bands according to [3].

Another problem was channel selection. All 16 channels gave different results. Channels 3, 6, 7, 9, 10, and 14 – 16 were most meticulously analysed. The results observed are patient dependent, including differences between electrode signals. This might have been caused by physiological differences and a slight difference in the location of electrodes caused by headcap displacements.

The power of the signal can change even for the selected band. One reason is the electrode's placement, while the second one can be due to balance itself. Even in the two feet standing position the signal can vary with time and this variation may be observed in the selected channels. Due to the very tight appliance of the headcap and the corresponding signals in neighbouring electrodes, it is unlikely that the cause are changes in contact impedance.

In the case of the one foot standing position, the signal variation can be observed as well. Some of the variations may be due to the brain's reaction to changes in the patient's position or tiredness caused by the length of the examination.

The problem consists in that in the event of losing balance the lead tension may change, which may cause short time variations in the measurement signal. This should be reduced by averaging power in a 60-second window (Fig. 2). Such averaging may also prevent the fast signal changes from being observed. This is why the comparison with a 1-second averaging window was carried out. In the case of the 1-second averaging window more rapid changes are visible, however, some of them may be due to distortion such as changes in lead tensions.

The influence of open/closed eyes on the signal registered was also observed. The difference between the one foot and two feet standing position was harder to observe for examinations carried out with eyes open. It was due to the role of the eyes in balance maintenance – it was harder for the patients to remain in an upright posture with the lack of visual stimuli (Figs. 3-4).

The main problem is that measurement data are not consistent. The results depend on the patient, channel and frequency band. However, in most cases there is a difference in signal power for different standing positions. These personal differences in the results obtained may be due to anatomical and physiological differences (especially the sense of balance).

The relation between power for two standing positions for the selected bands is person dependent (Figs. 5-7).

The most consistent results were obtained for closed eyes for channel number 6 and sigma frequency band.

The brain regions responsible for motor functions may also participate in the measured signals. This is why the analysis is so complicated.

According to the results obtained, new tests are prepared based on the evoked potentials (EP) approach. The plan assumes examination using periodic lifting of one foot.

V. Conclusions

The primary results show that it is possible to analyse the sense of balance using EEG, however, the frequency band and channel selection requires more investigation. Careful specification of the measurement protocol to remove unwanted sources of EEG signal also requires further verification – for example of conditions relating to open and closed eyes.

The problem of data analysis may follow from the complicated dependence of EEG signal on many sources like visual, motor aspects, etc.

The time variability of power courses for different time averaging window lengths should be also investigated as a possible indicator of the sense of balance.

It seems that further analysis for a larger number of people would allow to estimate the sense of balance on the basis of single channel or differential EEG measurements using time analysis for selected frequency bands.

References

[1] S. Chaudhuri, H. Thompson, and G. Demiris, "Fall Detection Devices and their Use with Older Adults: A Systematic Review", J Geriatr Phys Ther. 2014; 37(4): 178–196

[2] M. Kaczmarek, A. Bujnowski, J. Wtorek, A. Polinski, "Multimodal Platform for Continuous Monitoring of the Elderly and Disabled", Journal of Medical Imaging and Health Informatics, Volume 2, Number 1, March 2012, pp. 56-63(8)

[3] Y. Yi F. Tse, J. S. Petrofsky, L. Berk, N. Daher, E. Lohman, M. S. Laymon, P. Cavalcanti, "Postural sway and rhythmic electroencephalography analysis of cortical activation during eight balance training tasks", "Med. Sci. Monitor", 2013; 19:175-186

[4] Y. Ouchi, H. Okada, E. Yoshikawa, S. Nobezawa, M. Futatsubashi, Brain activation during maintenance of standing postures in humans. J Neurol, 1999; 122 (Pt 2): 329–38

[5] S. Dejardin, The clinical investigation of static and dynamic balance, B-ENT, 2008, 4, Suppl. 8, 29-36

[6] A. V. Oppenheim, G. C. Verghese, "Introduction to Communication, Control, and Signal Processing", Ch10, 2010

[7] http://www.biosemi.com/

[8] R. E. Challis, R. I. Kitney, Biomedical signal processing (in four parts). Part 3. The power spectrum and coherence function, Med Biol Eng Comput. 1991 May; 29(3):225-41

2nd International Workshop on Ubiquitous Home Healthcare

POPULATION aging is a phenomenon affecting many countries around the world. For example in Europe the life expectancy increased from 45 years in the early twentieth century, to 80 years now. Significantly longer life leads to age-related problems and diseases. In parallel, the cost of hospital care is increasing and additionally, a lack of the qualified caregivers is observed. Development of ubiquitous healthcare technologies can improve the quality of life of assisted citizens and can curtail growth in healthcare spending fueled by aging populations, and the prevalence of obesity, diabetes, cancer and chronic heart and lung diseases. In particular, information systems integrated with wearable, mobile devices and sensor networks at home can continuously assist persons while moving out or staying at home. Ubiquitous healthcare systems used as assisted living solutions will not only help to prevent, detect and monitor health conditions of a person but will also support of elderly, sick and disabled people in their independent living.

The goal of the UHH 2016 workshop is to gather researchers and engineers working in the field of ubiquitous healthcare to present and discuss new ideas, methods, and applications of assisted living IT technologies.

TOPICS

The workshop welcomes all work related to ubiquitous healthcare, but with a focus on the following themes (this list is not exhaustive):

- Ubiquitous healthcare information systems,
- Information processing algorithms for UHH,
- Ubiquitous services for home and mobile applications,
- Human-system interaction in UHH,
- Wearable sensors and systems,
- Smart glasses and smart watches in UHH,
- Data mining and knowledge discovery in ubiquitous healthcare,
- Integration of sensors and devices for UHH,
- Security of ubiquitous healthcare systems,
- Ensuring the Availability, Transparency, Seamlessness, Awareness, and Trustworthiness (A.T.S.A.T.) of home and mobile systems,
- Standardization in ubiquitous home healthcare,
- Applications of UHH for elderly, sick, and disabled people,
- Elderly care monitored dosage systems,
- Welfare technology for UHH.

The proposed papers should emphasize at least one of the following aspects:

- Assisted living,
- Home care,
- Self care,
- Mobile care.

BEST PAPER AWARD

During the closing ceremony the best paper award will be presented. Selection criteria will be based on results of the reviewing process and the quality of the presentation.

EVENT CHAIRS

- **Biallas, Martin,** iHomeLab, Hochschule Luzern, Switzerland
- **Wtorek, Jerzy,** Gdańsk University of Technology, Poland

PROGRAM COMMITTEE

- **Andrushevich, Alexey,** iHomeLab
- **Augustyniak, Piotr,** AGH University of Science and Technology
- **Bujnowski, Adam,** Gdansk University of Technology, Poland
- **Cavallo, Filippo,** Filippo Cavallo, The BioRobotics Institute, Scuola Superiore Sant'Anna
- **Haller, Michael,** University of Applied Sciences Upper Austria
- **Kaczmarek, Mariusz,** Gdansk University of Technology, Poland
- **Kistler, Rolf,** iHomeLab, Switzerland
- **Louventain, Nicolas,** SnT, University of Luxembourg
- **Martin, Benoit,** Université de Lorraine, France
- **McCall, Roderick,** Luxembourg Institute of Science and Technology
- **Pecci, Isabelle,** Université de Lorraine
- **Polinski, Artur,** Gdansk University of Technology, Poland
- **Popletev, Andrei,** SnT, University of Luxembourg
- **Rosell, Javier,** Universitat Politecnica de Catalunya
- **Ruminski, Jacek,** Gdansk University of Technology
- **Sincak, Peter,** Technical University of Kosice
- **Strumiłło, Paweł,** Lodz University of Technology
- **Svitek, Miroslav,** Czech Technical University in Prague
- **Tkacz, Ewaryst,** Silesian University of Technology, Poland
- **Truyen, Bart,** Vrije Univ. Brussels
- **Vogl, Anita,** MIL, University of Applied Sciences Upper Austria

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 247–254
DOI: 10.15439/2016F379
ACSIS, Vol. 9. ISSN 2300-5963

APIS – Agent Platform for Integration of Services

Michał Wójcik
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: michal.wojcik@eti.pg.gda.pl

Paweł Napieracz and Wojciech Jędruch
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: wjed@eti.pg.gda.pl

***Abstract*—This paper presents an approach to create a platform for development and evaluation of task execution algorithms relying on services composition. Proposed solution is based on an agent paradigm where autonomous agents can cooperate and negotiate in order to execute specified tasks which are defined by input/output descriptions. Tasks are realized by the means of services exposed by different agents. In case when there is no a single service fulfilling the submitted task requirements, there is a need for an automated composition of services into one complex workflow. The platform provides ready to use communication blocks which can be easily used for algorithms development without consideration for complex conversation protocols handling. All the algorithms developed on the platform are service implementation independent and oriented on inter-agent communication.**

I. Service Composition Problem

THE SERVICE composition is not a new problem and has been already considered in the literature. There is a number of different approaches to this problem:

- Centralized service providers systems which do not provide any kind of composition [1], but only give the user access to different resources in a uniform manner.
- Centralized systems providing workflow static composition and dynamic services selection features [2], [3]. Those require the user to define all the tasks in the workflow and the system itself performs only dynamic services selection based on the requested QoS parameters. They are often focused on the specific services architecture instead of generic algorithms.
- Decentralized agent service providers systems [4]. Those consider mainly broker agents providing features for discovering and negotiating services execution parameters. They do not provide any means of workflows composition but can be used as underlying systems.
- Decentralized agent systems with static workflow composition [5], [6]. In contrast to centralized systems, they can use autonomous agents for dynamic services selection. Agents acting as services brokers can negotiate compatibility and QoS parameters of the services.
- Systems, both centralized and decentralized, with dynamic workflow composition [7], [8]. Those require from the users only definition of the required output and optional input. Decentralized agent systems with dynamic composition are often used in the simulation of business processes used in the virtual organizations.

Most of the services composition systems are focused on the particular services architectures and description standards (UDDI, WSDL, OWL-S). The implemented cooperation algorithms are tested together with services efficiency which does not give the generic knowledge about the algorithms itself. Moreover, agent solutions not always consider agent communication standards which makes them even more limited to a particular solution.

A generic testbed environment, APIS (Agent Platform for Integration of Services), was created. Because the platform is not based on any particular service architecture it allows for testing cooperation algorithms with a focus on their performance rather than on services execution performance. The platform provides means for discovering, negotiating and executing abstract services using inter-agent communication based on the FIPA communication protocols [9].

There is a number of approaches for describing services for the composition process needs. Those can be full ontological descriptions concerning input/output syntactical definitions as well as semantical process definitions and some additional preconditions [10]. This allows for full ontological reasoning about services compatibility as well as desired output. Another known approach is syntactical input/output description combined with semantical service description based on thesaurus allowing for services matching based on words semantical distance [11]. Possibly simpler approach is semantical and syntactical input/output description with only I/O compatibility reasoning [12].

This paper proposes different solution, based on input/output and QoS ontological descriptions allowing for reasoning output $\rightarrow$ input compatibility between services. It assumes that I/O and QoS descriptions are enough for describing what and how should be done. Because this work does not consider different services architectures, at this point services adaptation has not been taken into consideration.

II. Agents as Service Providers

According to the most basic definition, an agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its delegated objectives. It might as well be an environment of some kind of services. Those can be both, Web Services distributed on remote machines connected to

the Internet as well as business services representing company activities mapped into computer system for the sake of simulations and automation. Agents can be treated both, as autonomous services providers and executors existing in such an environment. Moreover multi agent systems which assume communication and interaction between agents residing in the system, are suitable for this case.

When one agent is going to invoke a service of another one, there is a need for some kind of agreement between them to be established. Such an agreement should be made on the basis of some negotiations and be profitable for both sides. This actions can be described by Service Level Agreement (SLA) which is contractual obligations between a service consumer and a service provider, which can represent guarantees of quality of service (QoS), non-functional requirements of a service consumer and promises of a service provider [13].

III. Model of Self-Organization

This section presents a proposal of solution for the tasks composition problem. It distinguishes between different roles which can be taken by composing agents as well as between different communicative acts used in the composition process.

A. Agent Architecture

Figure 1 presents the APIS agent abstract architecture overview. It is a variation of the layered architecture where each of the layer can have a number of sub layers. All the layers (even the sub layers) can perceive input by a means of the see function which basically receives messages from other agents in the environment as well as produce output made of messages directed to those agents. Layers can be spawned dynamically by other layers and be attached as sub-layers or top level ones. All the layers are connected to agent inner state (for the sake of simplification, the architecture figure shows only one such a connection) which basically is a set of a services (both owned and those provided by other agents) known to the agent. Moreover, each of the layers contains its own state which is shared only with parent layer and sub layers. This state allows for performing long running actions based on previous interactions with other agents.

In this model, the only agent interactions with its environment are done through messages exchanged with other residing agents. The environment state can be described as one ore more messages (possibly from different agents) perceived within some context (e.g.: asking about particular service):

$$e = \{\mu_1, \mu_2, \ldots, \mu_n\} \tag{1}$$

where:

- $\mu \in \mathcal{M}$ which is a set of all possible messages.

This leads to defining agents actions also as set of messages (possibly addressed to different receivers):

$$\alpha = \{\mu_1, \mu_2, \ldots, \mu_n\} \tag{2}$$

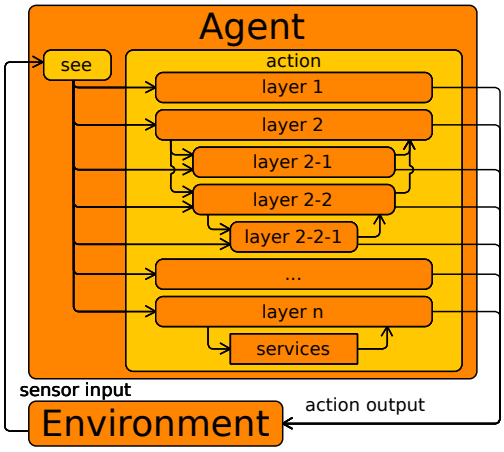


Fig. 1. Agent and its environment in the APIS platform

Finally, the agent decision function can be defined as a mapping from sequences[1] of messages sets[2] to messages sets:

$$a : \wp(\mathcal{M})^* \rightarrow \wp(\mathcal{M}) \tag{3}$$

The agent run function can be defined as subsequent environment state (message set) to action (messages set) transitions:

$$r : \mathcal{M}_0^i \xrightarrow{\mathcal{M}_0^o} \ldots \xrightarrow{\mathcal{M}_{u-1}^r} \mathcal{M}_u^o \tag{4}$$

where:

- $\mathcal{M}^i \in \wp(\mathcal{M})$ is a set of input messages,
- $\mathcal{M}^o \in \wp(\mathcal{M})$ is a set of output messages.

The run function can be formally presented as a mapping of environment states sequences and action sequences to environment states:

$$r : \mathcal{E}^* \times \mathcal{A}^* \rightarrow \mathcal{E} \tag{5}$$

Similarly to agent definition, the standard agent see function mapping environment state sequences to percepts can be defined as a mapping from a set of messages to percepts:

$$see : \wp(\mathcal{M})^* \rightarrow P \tag{6}$$

and the $action$ function mapping sequence of percepts to actions as a mapping from sequences of percepts to a set of messages:

$$action : P^* \rightarrow \wp(\mathcal{M}) \tag{7}$$

In layered architectures, decision function is realized through a set of behaviours, each associated with one layer. Because single layer can take part in ongoing inter agent negotiations, it can produce a number of different actions (messages sent to different agents) as well as be activated by a number of different environment states (messages from different agents) transformed into percepts. Because this is not a traditional layered approach where behaviours are described

[1]Sequences over set S are written as S^*
[2]Power set over set S is written as $\wp(S)$

as a pair of condition set and a resulting action there is a need for additional layer action function which defines how specified inputs are transformed into outputs:

$$beh = (P^c, \mathcal{A}^r, beh\_action) \tag{8}$$

where:

- P^c is set of percepts called the condition,
- $\mathcal{A}^r$ is set of possible actions called the result,
- beh_action is single layer action function.

A single behaviour action function can be defined as a mapping of percepts sequences and services sets to actions (sets of messages):

$$beh\_action : P^* \times \wp(Se) \to \wp(\mathcal{M}) \tag{9}$$

where:

- $Se = \{se_1, se_2, \ldots, se_{|S|}\}$ is a set of all services.

In order to compare different approaches to complex service workflows composition, a number of different utility functions can be introduced. A successful composition utility function returns values 1 and 0 determining if a composition process for a particular task was successful or not:

$$u_s : \mathcal{R} \to \{0, 1\} \tag{10}$$

where:

- $\mathcal{R} = \{r_1, r_2, \ldots, r_{|\mathcal{R}|}\}$ is a set of all possible runs in the environment.

A message utility function gives a natural number telling how many messages were used for a particular task execution:

$$u_m : \mathcal{R} \to \mathbb{N} \tag{11}$$

and a conversations utility function says how many different conversations have been started:

$$u_c : \mathcal{R} \to \mathbb{N} \tag{12}$$

A QoS utility function gives information about values of different QoS parameters describing the composed business process execution:

$$u_q : \mathcal{R} \to \wp(\mathbb{N}) \tag{13}$$

A time utility function tels how much real time (this value can vary on different hardware environment configurations) was used for the composition process:

$$u_t : \mathcal{R} \to \mathbb{R} \tag{14}$$

B. Agent Environment

The APIS platform assumes an agent system where a number of agents are spawned in order to cooperate. The multi-agent system can be formally described as:

$$sys = \langle A, env \rangle \tag{15}$$

where:

- A is a set of all agents in the system,
- $env = \langle \mathcal{E}, e_0, \tau \rangle$ is agent environment with initial state and state transfer function defined.

The classification of the agent environment in the APIS platform can be considered in two different scenarios concerning platform life time:

- short-run – all agents spawned at the same time only for single task execution request,
- long-run – agents can be spawned dynamically during platform lifetime, many independent task execution requests.

The environment characteristics which are common for both situations and does not change depending on platform life time are (based on [14, p. 30]):

- non-deterministic – because agents' actions consequences depend on inner states of all the agents taking part in a interaction, the single result can not be fully predicted,
- dynamic – dynamic environments change without agent interaction, the APIS platform environment can change only as a result of agents' action, but not all of the agents always take part in those interactions so the environment can change without their knowledge.

In the short-run, agents are spawned only for a single task execution and removed form platform afterwards. This method can be used for testing different algorithms and comparing them with different agents and services configurations. Moreover it can be used for cases when particular task should be carried on without any dependencies to other possible tasks. The environment characteristics in this situation are (based on [14, p. 30]):

- episodic – there is no connection between scenarios as agents are spawned only for a single task execution,
- discrete – there is finite number of environment and agents states (especially services composition possibilities) resulting from the initial platform configuration.

In the long-run, agents can be spawned and removed dynamically during the whole platform life-cycle. This allows agents to learn new composed services resulting from different tasks executions. This approach is especially useful in virtual organizations simulations. The environment characteristics in this situation are (based on [14, p. 30]):

- non-episodic – agents' decisions concerning the composition process are based on their knowledge about services, in a long-time running environment agents learn about new services and conditions negotiated at some point can influence future compositions,
- discrete – because number of agents residing on the platform, and services they know can change, there is an infinite number of services composition possibilities.

IV. INFRASTRUCTURE

The developed APIS platform [15], [16] is based on the JADE (Java Agent DEvelopment Framework) which is an agent development framework allowing for creating distributed multi agent systems [17]. It is one for the mostly used and recognizable agent platforms [5], [4], [3], [8], [6].

JADE supports behavior-oriented agent model, that means all agents actions are in a form of behaviors launched during

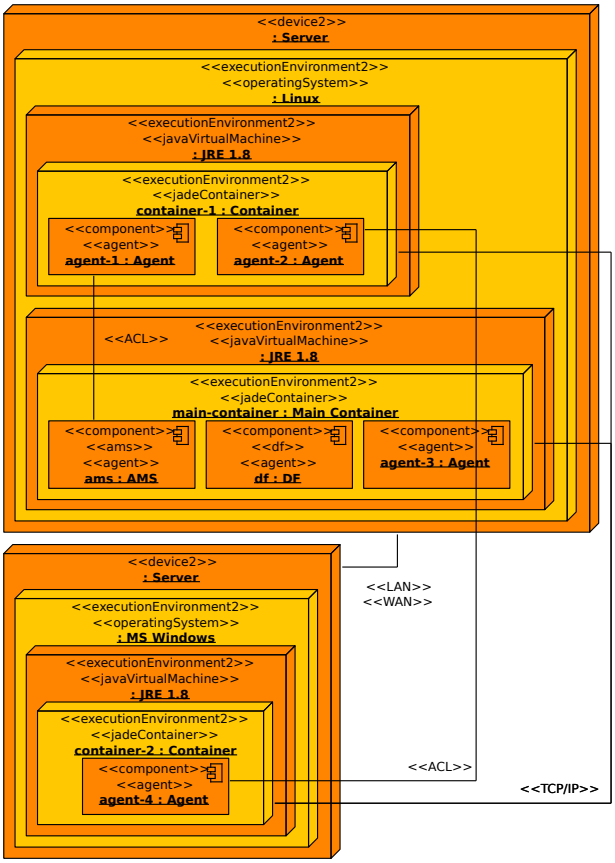


Fig. 2. Example JADE infrastructure

agent life cycle. One agent can make use of a number of different behaviors purposed for realization of different goals. New behaviors can be added while launching agent as well as during its life-cycle from other behaviors. This allows for dynamically adding behaviors related to the decisions made by an agent and clearly complies with the agent model defined in this paper.

An agent environment can be build with several nodes (in JADE called containers) which can run on one or more physical machines connected with network creating distributed environment. All the agents reside in those containers. JADE configuration requires at least one main container responsible for the whole platform, other dependent containers connect to the main one. JADE allows for configuration with a number of backup main containers synchronizing during life-cycle. Figure 2 shows an example JADE infrastructure. Only few connection between agents components were shown for the sake of simplification.

All agents belonging to the same platform can communicate with each other using ACL (Agent Communication Language), a standard language for agents communication defined by FIPA (The Foundation of Intelligent Physical Agents) [9]. Messages content is defined by two things: Semantic Language (SL) defining grammar of the message and domain ontology defining vocabulary. The language is defined by FIPA and is provided in JADE as a SL Codec whereas the ontology must be provided by the developer.

V. Agents Types

When concerning roles in the complex tasks execution process, different approaches can be taken. The APIS platform allows for deploying equal peers without any relations between them as well as agents organized into some hierarchies.

There are different roles, that can be taken by agents in the tasks execution process:

- client – an agent that searches for agents capable of executing particular task,
- principal – an agent that has some subordinates from whom it can request some actions,
- contractor – an agent able to expose some services, their QoS parameters as well as payment conditions,
- coordinator – additional role for contractor, introduced for better readability of centralized algorithms,
- subordinate – an agent that has a principal which can request some actions,
- collaborator – an agents collaborating with other agent on equal rights in order to perform some actions.

Then, different relationships between agents can be listed:

- client - contractor (coordinator) – client searches for contractors capable of executing particular tasks, during negotiations contractor provides QoS parameters which are evaluated by client,
- principal - subordinate – those are agents inside the same agency, where principal belongs to the sub-agency higher in the hierarchy, there are no negotiations and subordinate is not able to refuse performing requested tasks unless it is not capable of performing it,
- collaborator - collaborator – those can be agents in the same agency and both belonging to sub-agencies on the same level in the hierarchy, they can be requested by principal to perform some tasks together and they must jointly work out a solution.

One agent can be in more than one role at the same time. For example the same agent can be contractor for external client, collaborator for agents in the same sub-agency and principal for agents in sub-agencies lower in the hierarchy at the same time.

VI. Communication Protocols

According to the FIPA standard, there is a number of protocols describing in details communication between agents [9]. In order to ensure that the APIS platform is complying with the FIPA standard, all the composition algorithms should be based on the FIPA protocols. Initially, for the needs of this work the five protocols were chosen: cancel, request, query, contract net and iterated contract net protocols.

The cancel meta protocol allows the initiator to cancel on going interaction with another participant under any protocol [9]. The initiator trying to cancel an interaction, needs to send `cancel` message containing the message that it

wants to cancel. The participant can reply with `inform` message after successful termination or `failure` message when termination did not succeed.

The query protocol allows the initiator to ask the participant if a given proposition is true by sending `query-if` message or ask for information concerning a given object by sending `query-ref` message [9]. The participant can agree or not agree to respond by replying with respectively `agree` or `refuse` message. If the participant agreed to response, it sends `inform` message containing true/false reply or information concerning provided object. It can also send `failure` message if an attempt to acquire the answer finished with an error.

The request protocol allows the initiator to request an execution of a given action by the participant by sending a `request` message [9]. The participant can refuse performing the given action by replying with a `refuse` message or agree by replying with an `agree` message. The participant replies with an `inform` message which contains an action outcome. The participant can reply with a `failure` message if performing the action finished with an error.

The contract net protocol allows the initiator to gather a number of proposals of performing some action from one ore more participants [9]. Firstly, the initiator sends a `cfp` message containing action description to potential performers. The participants can reply with a `propose` message containing some conditions of executing the given action or with a `refuse` message when they are not interested. After gathering all the replies or exceeding a deadline (specified in the first `cfp` message) the initiator browses the proposals and selects one or more the best ones. Authors of the selected proposals receive an `accept-proposal` message and authors of the rejected ones receive a `reject-proposal` message. All the proposals received after the given deadline are automatically rejected and their authors receive a `reject-proposal` message with a corresponding information. Similarly to the request protocol, after finishing the action, the participants send an `inform` message which can contain an action result or `failure` message in case of a failure.

The iterated contract net enriches the contract net protocol with a possibility of stating more exact conditions in a negotiation process [9]. After selecting propositions, the initiator can decide if that was or not the final iteration. If it was the final iteration, protocol proceeds as in standard contract net protocol. If it was not the final iteration, the initiator sends more exact `cfp` message. This process can be repeated until the initiator decided, that further negotiations are not required.

VII. Algorithms Building Blocks

The idea of the APIS platform is to develop complex tasks execution algorithms using pre-made block providing all inter-agent communication actions so the developer can focus on the algorithms structure. Normally developer would be forced to implement all the communication stack including packing, unpacking, sending, receiving and filtering messages within a number of ongoing conversations.

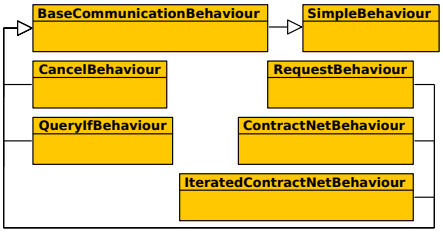


Fig. 3. Active behaviours

Figure 3 presents hierarchy of behaviours which can be used by the developer while creating an active side of the execution process. Those behaviours are:

- `SimpleBehaviour` – basic JADE behaviour class for all agent behaviours,
- `BaseCommunicationBehavior` – basic communication stack operations common for all the used protocols,
- `CancelBehaviour` – implementation of *meta cancel* protocol, allows to cancel any ongoing conversation,
- `QueryIfBehaviour` – implementation of *query if* protocol, allows to check if given fact is true according to other agents, requires only providing the fact and a receiver,
- `RequestBehaviour` – implementation of the *request* protocol, allows to request performing some action by another agent, requires providing the action definition and a receiver,
- `ContractNetBehaviour` – implementation of the *contract net* protocol, allows to call for proposals of performing some action, gather those proposal, select the best one and gather the result, requires providing the action definition, a list of receivers and a proposals comparator,
- `IteratedContractNetBehaviour` – implementation of *iterated contract net* protocol, allows to do the same as the `ContractNetBehaviour` but with negotiation iterations.

Figure 4 presents hierarchy of behaviours which are automatically used by the passive side of the execution process. Those can not be used directly by the developer and are launched automatically by the agent when specified initiating messages is received. Those behaviours are:

- `ResolverBehaviour` – basic resolving behaviour implementing common communication stack,
- `CancelResolverBehaviour` – behaviour launched when a *cancel* message is received,
- `QueryIfResolverBehaviour` – behaviour launched when a *query if* message is received, checks submitted fact and responds with a result,
- `RequestResolverBehaviour` – behaviour launched when a *request* message is received, performs requested action and responds with a result,
- `CallForProposalResolverBehaviour` – behaviour launched when a *call for proposal* message

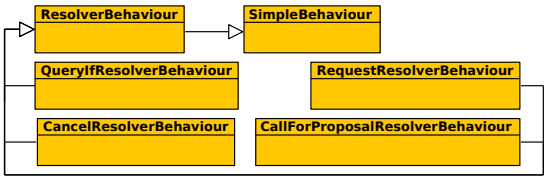


Fig. 4. Passive behaviours

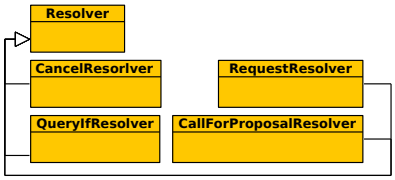


Fig. 5. Passive behaviours resolvers

is received, prepares proposal and if accepted performs request action and responds with a result.

While the passive behaviours are fixed, their outputs can be changed by the mean of resolvers. The resolvers are interfaces which can be implemented by the developer and which are used by the passive behaviours. Figure 5 presents a hierarchy of resolvers used on the platform. Those interfaces are:

- `Resolver` – basic interface for all the resolvers,
- `CancelResolver` – defines actions carried on by the agent after receiving *cancel* message,
- `QueryIfResolver` – defines actions for checking if giver fact is true,
- `RequestResolver` – defines actions of executing specified action and returning its result,
- `CallForProposalResolver` – defines action of preparing proposal and if accepted execution specified action and returning its result.

VIII. Simple Composition Algorithm

In order to show that the platforms fulfills its requirements, the simple composition algorithm was prepared. During its implementation no communication based code was prepared. Figure 6 presents classes which were developed and their relation to those provided by the platform. Those classes are:

- `Coordinator` – agent coordinating the composition process, registers appropriate request resolver,
- `Contractor` – agent providing some services, registers appropriate call for proposal resolver,
- `ReqResolver` – implementation of the request resolver, when receiving a request to execute some task it starts the composition behaviour,
- `CFPResolver` – implementation of the call for proposal resolver, checks if requested task output can be provided by any of the services known by the resolver owner and if yes prepares an appropriate proposal,
- `CompositionBehaviour` – a behaviour responsible for composing a new workflow of services in order to provided desired task output, in order to find subsequent

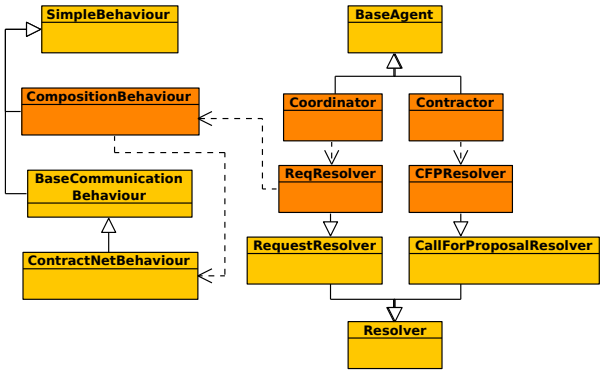


Fig. 6. Simple algorithm

workflow services it starts underlying contract net behaviours.

IX. Running the Algorithm

The designed algorithm was carried on by the following agents exposing specified services:

- agent-pizza-maker:
  - MakePizza : (base, topping, sauce) → (pizza);
- agent-baker:
  - MakeBase : (flour, water) → (base);
- agent-sauce-maker:
  - MakeSauce : (tomato, water) → (sauce);
- agent-topping-maker:
  - MakeTopping : (vegetable) → (topping);
- agent-seller:
  - ProvideFlour : () → (flour),
  - ProvideWater : () → (water),
  - ProvideVegetable : () → (vegetable),
  - ProvideTomato : () → (tomato);
- agent-coordinator,
- testRunner.

All the agents with the *agent-* prefix are contractors without any hierarchical relationships. The agent-coordinator is an agent which receives a request from the client (the testRunner agent). Only services exposed by the agent-seller agent do not require any input so they should be used as the workflow initial services.

The subsequent messages exchanged by the agent are presented in figures from 8 to 11 which were created with APIS version of JADE sniffer agent. The communication snapshot presents which messages belong to which conversation. The explanation of goals of each conversation is presented in table I. The final workflow providing desired output is presented in figure 7 which was created using the APIS service sniffing tool. Values for the utility (10, 11, 12, 13, 14) functions are: $u_s = 1$, $u_m = 129$, $u_c = 55$, $u_q = 9$, $u_t = 150.5$ms.

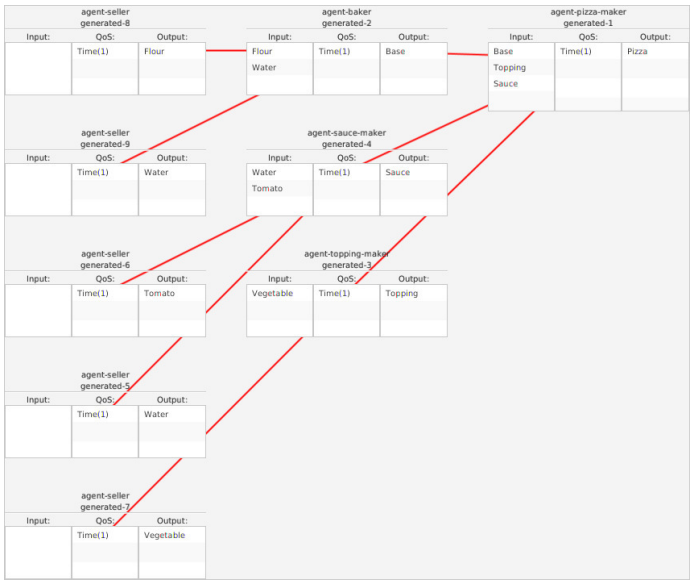


Fig. 7. Service *generated-0* chain for simple centralized algorithm

TABLE I
CONVERSATIONS SUMMARY FOR SIMPLE CENTRALIZED ALGORITHM

conversations	protocol	output	service
0	request	pizza	generated-0
1, 7, 13, 19, 25, 31, 37, 43, 49	request	agents	-
2, 3, 4, 5, 6	cfp	pizza	generated-1
8, 9, 10, 11, 12	cfp	base	generated-2
14, 15, 16, 17, 18	cfp	topping	generated-3
20, 21, 22, 23, 24	cfp	sauce	generated-4
26, 27, 28, 29, 30	cfp	water	generated-5
32, 33, 34, 35, 36	cfp	tomato	generated-6
38, 39, 40, 41, 42	cfp	vegetable	generated-7
44, 45, 46, 47, 48	cfp	flour	generated-8
50, 51, 52, 53, 54	cfp	water	generated-9

X. CONCLUSION

The idea of this work was to provide an agent platform allowing for developing and evaluating complex tasks execution and services composition algorithms. In order to focus on the algorithms, the platform is not based on any services implementation but only on inter agent FIPA communication standard. It has been shown how APIS algorithms building blocks comply with the FIPA communication protocols and that they can be successfully used in developing execution and composition algorithm. The set of building blocks can be easily expanded with new protocols by implementing low-level communication stack. Moreover it has been shown that platform accompanying sniffing tools allow for good algorithms evaluation. Proposed utility functions can be used for comparing different algorithms.

Despite the fact that the APIS platform is not based on any services implementation it can be in future easily enriched with one by changing appropriate algorithms building blocks. As future work, the most important possibility is developing and evaluating more complex execution and composition algorithms, especially distributed ones based on work division.

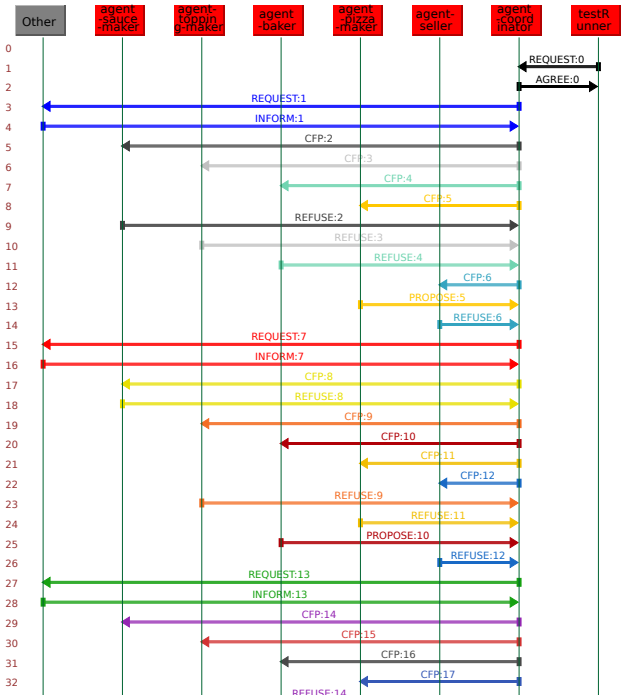


Fig. 8. Communication snapshot

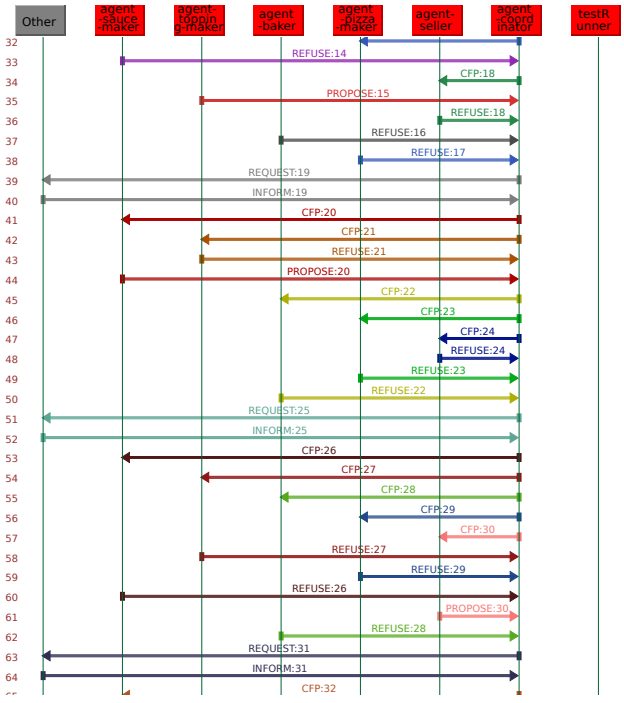


Fig. 9. Communication snapshot (continued)

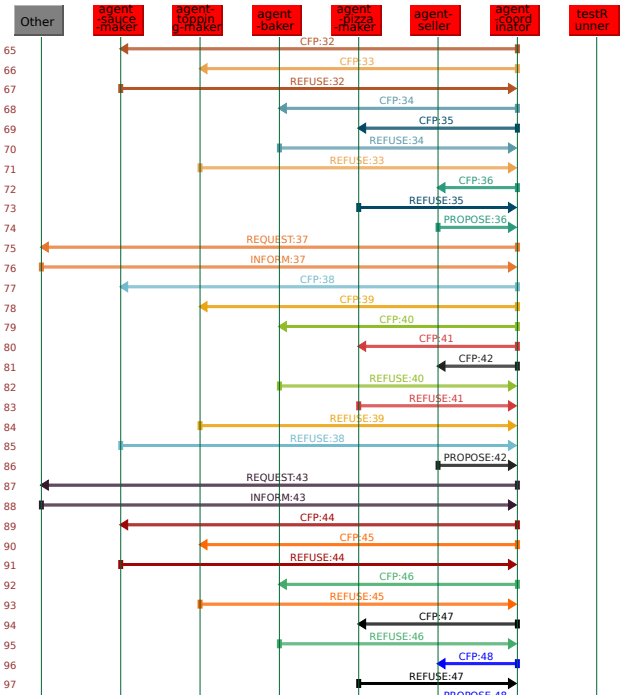


Fig. 10. Communication snapshot (continued)

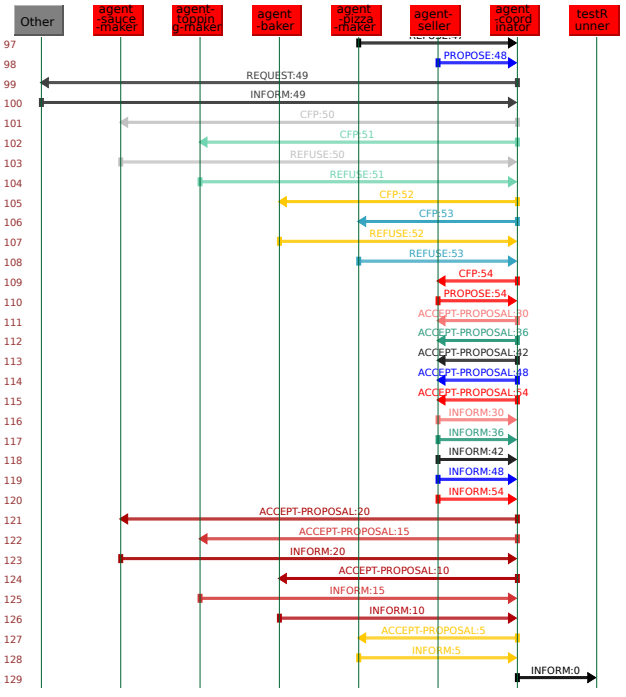


Fig. 11. Communication snapshot (continued)

REFERENCES

[1] P. Czarnul, "A JEE-Based Modelling and Execution Environment for Workflow Applications with Just-in-Time Service Selection," in *Proceedings of the 2009 Workshops at the Grid and Pervasive Computing Conference (GPC)*, ser. GPC '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 50–57. [Online]. Available: http://dx.doi.org/10.1109/GPC.2009.24

[2] ——, "Comparison of selected algorithms for scheduling workflow applications with dynamically changing service availability," *Journal of Zhejiang University SCIENCE C*, vol. 15, no. 6, pp. 401–422, 2014. [Online]. Available: http://dx.doi.org/10.1631/jzus.C1300270

[3] F.-S. Hsieh and J.-B. Lin, "Context-aware workflow management for virtual enterprises based on coordination of agents," *Journal of Intelligent Manufacturing*, vol. 25, no. 3, pp. 393–412, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10845-012-0688-8

[4] L. Ehrler, M. Fleurke, M. Purvis, B. Tony, and R. Savarimuthu, "AgentBased Workflow Management Systems (WfMSs): JBees - A Distributed and Adaptive WfMS with Monitoring and Controlling Capabilities," in *Journal of Information Systems and e-Business Management, Volume 4, Issue 1*. Springer-Verlag, 2005, pp. 5–23. [Online]. Available: http://dx.doi.org/10.1007/s10257-005-0010-9

[5] P. Czarnul, M. Matuszek, M. Wójcik, and K. Zalewski, "Beesybees: A mobile agent-based middleware for a reliable and secure execution of service-based workflow applications in beesycluster," in *Multiagent and Grid Systems*. IOS Press, 2011, vol. 7, pp. 219 – 241. [Online]. Available: http://dx.doi.org/10.3233/MGS-2011-0178

[6] P. Czarnul and M. Wójcik, "Dynamic compatibility matching of services for distributed workflow execution," in *Parallel Processing and Applied Mathematics*, ser. Lecture Notes in Computer Science, R. Wyrzykowski, J. Dongarra, K. Karczewski, and J. Wasniewski, Eds. Springer Berlin / Heidelberg, 2012, vol. 7204, pp. 151–160. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31500-8_16

[7] F. E. Tosta, V. Braganholo, L. Murta, and M. Mattoso, "Improving workflow design by mining reusable tasks," *Journal of the Brazilian Computer Society*, vol. 21, no. 1, pp. 1–16, 2015. [Online]. Available: http://dx.doi.org/10.1186/s13173-015-0035-y

[8] F.-S. Hsieh and J.-B. Lin, "A self-adaptation scheme for workflow management in multi-agent systems," *Journal of Intelligent Manufacturing*, vol. 27, no. 1, pp. 131–148, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10845-013-0818-y

[9] The Fundation of Intelligent Physical Agents, "FIPA specifications," Tech. Rep., 2002. [Online]. Available: http://www.fipa.org/repository/standardspecs.html

[10] K. Sycara, M. Paolucci, A. Ankolekar, and N. Srinivasan, "Automated discovery, interaction and composition of semantic web services," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, no. 1, pp. 27 – 46, 2003. [Online]. Available: http://dx.doi.org/10.1016/j.websem.2003.07.002

[11] K. Arisha, F. Ozcan, R. Ross, S. Kraus, and V. S. Subrahmanian, "Impact: the interactive maryland platform for agents collaborating together," in *Multi Agent Systems, 1998. Proceedings. International Conference on*, Jul 1998, pp. 385–386. [Online]. Available: http://dx.doi.org/10.1109/ICMAS.1998.699225

[12] G. Wickler and A. Tate, "Capability representations for brokering: A survey," in *Available from: www.aiai.ed.ac.uk/ âĹij oplan/cdl/cdl-ker.ps*, 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.367.9865

[13] Q. He, J. Yan, R. Kowalczyk, H. Jin, and Y. Yang, "Lifetime service level agreement management with autonomous agents for services provision," *Inf. Sci.*, vol. 179, no. 15, pp. 2591–2605, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2009.01.037

[14] S. J. Russell and P. Norvig, *Artificial Intelligence a modern approach*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 2003.

[15] P. Napieracz, "Porównanie agentowych algorytmów kooperacji w wykonywaniu złożonych zadań," Master's thesis, Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, 2014.

[16] M. Wójcik, "Raport techniczny nr 2/2015: Projekt platformy apis (agent platform for integration of services)," Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Tech. Rep., 2015.

[17] F. L. Bellifemine, G. Caire, and D. Greenwood, *Developing Multi-Agent Systems with JADE*. Wiley, 2007.

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 2016 are:

- MAS&S'16 - 10^{th} International Workshop on Multi-Agent Systems and Simulations
- SEN-MAS'16 - 4^{th} International Workshop on Smart Energy Networks & Multi-Agent Systems

10^{th} International Workshop on Multi-Agent Systems and Simulations

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'16 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

STEERING COMMITTEE

- **Cossentino, Massimo,** ICAR-CNR, Italy
- **Fortino, Giancarlo,** Universita della Calabria, Italy
- **Gleizes, Marie-Pierre,** Universite Paul Sabatier, France
- **Pavon, Juan,** Universidad Complutense de Madrid, Spain
- **Russo, Wilma,** Universita della Calabria, Italy

EVENT CHAIRS

- **Fortino, Giancarlo,** Universita della Calabria, Italy
- **Fuentes-Fernández, Rubén,** Research Group on Agent-based, Social & Interdisciplinary Applications (GRASIA), University Complutense of Madrid (UCM), Spain
- **Niazi, Muaz,** COMSATS Institute of IT, Pakistan
- **Seidita, Valeria,** Università degli Studi di Palermo, Italy

PROGRAM COMMITTEE

- **Alam, Shah Jamal**
- **Antunes, Luis**
- **Arcangeli, Jean-Paul,** Université Paul Sabatier, France
- **Bernon, Carole,** Universite Paul Sabatier, France
- **Cipresso, Pietro**
- **Cossentino, Massimo,** ICAR-CNR, Italy
- **Davidsson, Paul,** Malmö University, Sweden
- **Garro, Alfredo,** University of Calabria, Italy
- **Gomez-Sanz, Jorge J.,** Universidad Complutense de Madrid, Spain
- **Gravina, Raffaele,** University of Calabria, Italy
- **Guerrieri, Antonio,** University of Calabria, Italy
- **Klügl, Franziska,** Örebro Universitet, Sweden
- **Molesini, Ambra,** Università di Bologna, Italy
- **Petta, Paolo,** OFAI, Austria
- **Ribino, Patrizia,** Istituto di Reti e Calcolo ad Alte Prestazioni - Consiglio Nazionale delle Ricerche, Italy
- **Savaglio, Claudio,** Universita della Calabria
- **Vizzari, Giuseppe,** Università di Milano Bicocca, Italy

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 259–263
DOI: 10.15439/2016F578
ACSIS, Vol. 9. ISSN 2300-5963

Talents, Competencies and Techniques of Business Analyst: A Balanced Professional Development Program

Anna Bobkowska
Faculty of Electronics, Telecommunications
and Informatics, Gdansk University of Technology
Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: annab@eti.pg.gda.pl

***Abstract*—This paper presents preliminary results of action research in which we search for fundamentals of an universal theory of balanced approaches to software process. It is developed on the basis of balanced approach for professional development program for business analysts which integrates approaches oriented on talents, competencies and techniques. This paper includes the description of key concepts in background approaches, components of the framework for professional development program for business analysts and preliminary results of the universal theory of balanced approaches to software process.**

I. Introduction

WHEN asking the question "Can everyone become a good business analyst?" during the trainings for students and professionals, nearly 100% of answers is "no". The role of business analysts in software projects becomes increasingly important and challenging [1-3]. Business analysts perform their tasks on the edge of several domains. They are expected to provide a link between business and technology. They should support managers in making decisions about innovations. The results of their performance have a large impact on the work of software developers and the final success or failure of software projects. Tasks of business analysts require huge knowledge, several behavioral skills, proper attitudes, and probably… real talents. Additionally, it seems that business analysts have to learn continuously and acquire diversified experience in order to deal with more and more advanced challenges. Thus, the problem of effective professional development program is important to individual business analysts who want to develop their career. Additionally, it is in scope of interest of coaches, mentors or tutors who can support business analysts in their professional development as well as project managers or human resources (HR) managers who are responsible for goal achievement in projects or organization capability.

Traditional approaches to education of system or business analysts are based on teaching several techniques, including requirements engineering, software modeling and business modeling. On the other hand, best practice rooted in industry emphasizes the importance of behavioral competencies, such as communication or interaction skills. Since they depend on personality features and they are difficult to assess by teachers, they often are skipped in education. Talent-oriented or career-oriented approaches are not popular in business analysis education, with a few exceptions [4-5]. The term of `talent` appears mainly in job offers for business analysts, but not much research has been done in this area. Approaches based on methodology, competencies and talents have different fundamental assumptions. Thus the following questions can be posed: Are these approaches contradictory or complementary? Is there another kind of relationship between them? How to integrate them? How to balance actions recommended by them? Could a universal theory of balanced approaches be helpful in performing these tasks?

The goals of the paper are:

- to present a balanced approach to professional development of business analysts which integrates talent-oriented approach, competence development and methodological software process definition,
- to search for fundamentals of an universal theory of balanced approaches to software process.

We apply the research methodology inspired by action research [6]. When solving a practical problem of supporting business analysts in their professional development, we attempt to find universal concepts of balanced approaches which can be useful also for solving the problems of balanced approaches in other contexts. In our argumentation, we use also other cases of balanced or integrated approaches.

The paper is structured as follows. Section 2 presents key concepts in the background approaches. Section 3 describes components of the framework for balanced approach to the program of professional development of business analysts. Section 4 presents preliminary results of searching for fundamentals of the universal theory of balanced approaches. Section 5 draws conclusions and presents prospects for further work.

II. Key Concepts in Background Approaches

The background approaches take different perspectives which is manifested by different terminology and patterns of action. In order to be precise in our considerations, we select a few characteristic approaches and mark their `key concepts`. We also attempt to make `connection points` between them with the purpose of preparation to a balanced approach.

The following background approaches are used:

- methodological approach represented by IBM Rational Unified Process (IBM RUP) which has defined a group

of analysts roles in software development process and has assigned to them tasks and work products [7-9],
- competencies described in "A Guide to the Business Analysis Body of Knowledge (IIBA BABOK Guide) [2],
- talent-oriented approach [10, 11] which is based on recognition of personal talents and continuous development of professional competencies.

A. Methods and Techniques

IBM RUP defines roles in software development process. The group of analysts includes the following roles: business architect, business designer, business-process analyst, requirements specifier, stakeholder and system analyst. The roles have very precise descriptions. They are related to tasks and to work products for which the roles are responsible. A set or required skills is defined for each role. For example, "a person acting as business-process analyst must be a good facilitator and have excellent communication skills; knowledge of the business domain is essential for those acting in this role". Roles, tasks and work products are used in configurations of software process for given types of projects and they are used for preparing plans of software development for concrete projects. In context of company, analysts are expected to perform their tasks according to their role description in order to produce work products. They can use several kinds of guidelines. The main values are methodological rigor and efficient performance of high quality work products.

IBM RUP contains description of skills, which makes a connection point to competence-based approach. The detailed description of roles can be used for career path definition. However, when changing the company, a person can face other career paths. A set of methodological role definitions may direct individual program of competence development by indicating the skills which are important in a given professional situation.

B. Business Analyst Competencies

IIBA BABOK Guide defines a business analyst as "any person who performs business analysis tasks described in the *BABOK® Guide*, no matter their job title or organizational role." For effective performance, business analysts should have competencies in the following groups:
- Analytical thinking and problems solving;
- Behavioral characteristics, such as ethics, trustworthiness, and personal organization;
- Business knowledge about general principles, industry-specific, organizational and knowledge about solutions;
- Communication skills;
- Interaction skills including facilitation and negotiation, leadership and teamwork; and
- Tools and technology.

Effectiveness measures are described for all competencies which allows for making assessments. The main value of this approach is in adaptability to the needs based on professional skills. Techniques in IIBA BABOK guide constitute a kind of toolbox.

The extensive description of underlying competencies makes a significant contribution of IIBA, but the IIBA BABOK Guide contains also methodological components such as knowledge areas, tasks in the knowledge areas and related techniques.

C. Talent-Oriented Approach

A popular way of thinking about changes in personal development is a transition from a present (AS-IS) state to the desired (TO-BE) state. When this approach is integrated with talent recognition, it might lead to perfection. The elements of the change can include the knowledge, the practical use of some techniques and the behavioral skills and attitudes.

Talent management is one of the trends in HR management nowadays. Talents are naturally recurring patterns of thought, feeling or behavior that can be productively applied to achieve a nearly-perfect performance of a given activity. The activity is performed with ease and pleasure. It gives satisfaction. On contrary, the lack of talent causes several difficulties and negative emotional attitudes when performing the activity. Talent management can be considered from both individual and organizational perspectives. There are several topics in debate about talents, including the percentage of talented employees in organization ranging from 2% to entire 100%; the features which constitute talents; possible dependence on organizational context; and effective ways of talent management. It is believed that talent, passion and hard work are key factors of every success.

Talented individuals who want to become successful business analysts should individually elaborate vision of their careers and find ways of making progress by gaining experience and competence development. The connection point to methodologies is the following: talent development can be supported by taking a sequence of more and more challenging positions or roles. Explicit knowledge of required competencies can help in conscious improvement of their skills.

III. Balancing Talents, Competencies and Techniques in Professional Development Program

This section describes an attempt to integrate approaches based on talents, competencies and methodologies of business analysts in a framework for balanced professional development program.

A. Framework Overview

The main components of the framework are presented in Fig. 1. The fundamental element of professional development program is a change from assessment of present (AS-IS) state to the vision of desired (TO-BE) state. Obviously not just one change, but several changes should be performed iteratively. For the first change, a pure self-assessment and a rough vision of the desired TO-BE state should be made. Starting from the second iteration, results of action from previous iterations can be used as criteria of assessment and progress evaluation. In further steps, effectiveness measures find their application.

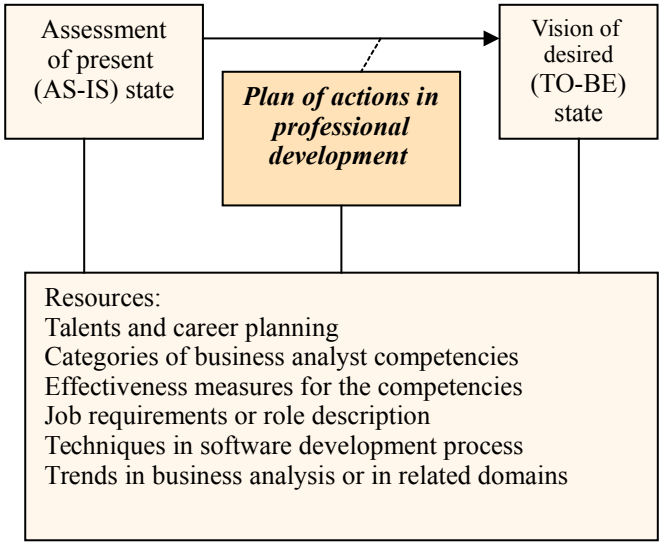


Fig. 1. A framework for professional development program

The individual program includes also `plan of actions` on the way of making the change. It can be related to gaining new knowledge about principles of business, certain industry or organization. It can contain a task to learn a new business analysis technique or CASE tool features. The training can be achieved by self-studies or participating in specialized courses. In other cases, skills can be acquired by experience of performing certain kinds of activities. Probably, the most difficult to gain are the behavioral competencies, such as communication skills and interaction skills. However, with the progress in the area of personal development training and coaching, these kinds of change are also possible.

B. Resources

Both the vision of TO-BE state and the assessment of AS-IS state can be supported by resources which can constitute a frame of reference for making proper decisions.

The change can be based on recognition of talents and ambition of career development. In this case, one can start with their strengths, practice related activities and develop complementary skills which allow for solving more and more advanced problems.

The change can be driven by career path defined in a company. In this case, it starts with a list of job requirements or role description in software process (e.g. role definitions in companies which develop software according to IBM RUP). The training can facilitate further steps in career path. The job requirements can be also used as a checklist for self-assessment and serve as a basis for formulation of the vision of TO-BE state.

When specific job requirements are not available or skill development is made with a more general purpose, one can use more general frames of reference, such as categories of business analyst competencies in IIBA BABOK Guide. In this case, detailed effectiveness measures can be useful for tracing the progress.

For advanced business analysts, professional literature on business analysis trends or advances in business knowledge in related domains can be used in continuous professional development.

C. Discussion

When attempting to reflect on the framework from meta-conceptual perspective, one can notice that the core concept of change has been taken from personal development perspective. This is because it is the most closely related to the problem to be solved. Other background approaches contribute to resources in an additive manner. It is good to see their unified role. But on the other hand, there are no recommendations how to balance use of resources in given circumstances. In order to capture this issue more precisely, some knowledge about general rules of balanced approaches would be useful.

This framework has been verified in a case study. It has shown usefulness of the terminology for description of real case and for directing iterative professional development program. Additionally, it highlighted aspects related to practical application, such as the role of coach in this process.

IV. Towards a Universal Theory of Balanced Approaches

A. Universal Problem

The concept of balanced approaches to software process has a much broader scope. Almost all software companies face the challenge of changes in the surrounding world which force them to adapt. They have usually worked out some effective software processes related to their expertise. On the other hand, new technologies and approaches are being proposed. How to keep their best practice and improve with new approaches? How to keep balance in this activity? The best example of common challenge which several companies had to face recently was the confrontation of methodological software development processes with the trend of agile approaches. Some companies had to deal also with increasing requirements regarding usability and user experience. Other trendy approaches include: business process management, collaborative software development, mobile application, gamification, smart technologies, idea of cashless world, etc. One can expect, that novel approaches are being continuously proposed in future. In consequence, software companies will have to address the challenges of innovation and adaptation continuously. Can we facilitate it with an universal theory of balanced approaches?

B. From Balanced Program to Universal Theory

Having the results of balanced professional development program with a case study of its application in practice, we are going to see how it can contribute to a unified theory of balanced approaches. The second question is whether they are general or not.

We have started with key concepts of approaches to be balanced and we were looking for connection points. The problem of balancing any approaches is not about providing solutions from scratch. It is always based on some background approaches. Furthermore, we need to integrate them before we can balance them.

Several strategies were used in order to integrate and balance the approaches. Selection of the core approach of change in professional development program and collecting resources from all background approaches in an additive manner to form a frame of reference were performed at the stage of building the framework. Selection of the first-need skill, best-fit techniques, or continuous grow strategies were used in practical application. It seems they were chosen based on changing circumstances. Factors considered for background approaches included: motivation, terminology, context of discovery, strengths and weaknesses. Thus, there is not just one proper way of balancing several approaches. In general, the diversity of possible ways how to integrate and balance can be captured by patterns.

The case study has shown that not all aspects of practical application were considered in the framework, especially the activity of coach. In general, problems addressed by balanced approach usually have practical applications. Thus, the need to provide both theory and recommendations for practical use seem to be universal to several cases.

C. Patterns of Balanced Solutions

The universal theory of balanced approaches can provide patterns of relationships between aspects which have their background in different approaches. The assumption is that a few background approaches are valuable, i.e. these are not cases of selecting one with rejection of others, or replacing older technology with a newer one.

The following `patterns` are possible:

- Independent dimensions: background approaches constitute different dimensions of software process and one can work independently in any of the dimensions; the only question is about priorities,
- Dimensions with overlapping: background approaches constitute different dimensions of the process with some overlapping, thus some connection points should be defined; e.g. [12],
- Time-ordered balance: different background approaches prevail in different stages of software development process, e.g. intensive use of creativity techniques during the analysis of innovative products and methodological software development in the following phases [13],
- Core approach with elements of other approaches: selection of the core background approach with incorporation of elements of other approaches taking into account a given criterion, e.g. balance between agile and disciplined software development with risk management [14],
- Process innovation: incorporation of novel approaches into existing process, e.g. experiments with gamification approaches when keeping application of best practice of software development process,
- Application dependent on circumstances: when several background approaches have different area of application or a different context of discovery, the decision can be made on the basis of analysis of the circumstances of the problem at hand and the fundamentals of the background approaches.

D. Other Components of the Balanced Approach

A logical consequence of the discovery that "it depends on several factors" is the question "what are these `factors`?" In search of them, one can use typical characteristics of software projects, such as type of product under development, complexity of software project, culture of company, stakeholders, team experience, main risks, legal and formal requirements, etc.

Another interesting question is whether we can use knowledge from other `disciplines` in order to avoid developing this theory from scratch in situation when several research results could be used by analogy or by knowledge transfer. What are these disciplines? What is specifics of their application? Innovation management seems to be a good source of knowledge on innovations in industrial context. Areas of multi-dimensional analysis and optimization can support increase of the level of methodological precision. Research results of empirical software engineering can deliver several interesting finding regarding the performance of software engineering processes in diversified situations. Identification of other disciplines acquiring insights from them is the task for further studies.

V. Conclusions and Further work

This paper presents work in progress and, as such, it generates more questions than it can answer. The addressed problems seem to be universal. If proper solutions are found, they can facilitate dealing with several situations of balancing different approaches in practice.

Regarding the first goal, this paper has presented the balanced approach to individual program of professional development for business analysts. The result is the framework which includes self-assessment of talents and competencies in confrontation with job requirements and methodologies. It should be performed iteratively as the AS-IS state changes and business analyst become more aware of their talents as well as barriers in professional development. Business analysts might also face new challenges or they might aim at perfection in their professional development with use of efficiency measures for the competencies they work on as well as professional literature related to trends in business analysis.

This approach allows for conscious direction of professional development activities. In contrast to most of available professional trainings, its fundamental assumptions are individual programs and continues training. This is an important issue in times when everyone can expect several job changes, life-long learning is an increasing trend and personal responsibility for the professional career is reality.

Integration and balance appear to be beneficial to the professional development program. The approaches were based on complementary values. Talent-oriented approach focused on individual side of the professional development. Methodologies describe best practice of successful software development including placement in software development process and templates of work products. Competencies are like a link between techniques and personal talents. They

allow business analysts to become aware of their talents and to make verified steps towards excellence.

There are several issues which should be added to the framework. For example, development of effective methods of training on skills related to specific business analyst behaviors are a challenge. Empirical studies from several cases of career development could provide validation of the proposed approach and many valuable insights.

Regarding the second goal, in this first step towards the universal theory of balanced approaches, several patterns of balance have been identified. Other components include: ways of integrating the background approaches before they can be balanced, consideration on both levels of theory and pragmatics, a set of factors, on which a pattern and/or decision depends. The factors include (but are not limited to) level of advancement in performing balanced activities, characteristics of software projects, comparison of the context of discovery and the context of application of a given approach. It is expected that knowledge from several disciplines can be applied by analogy or by knowledge transfer, but the complete list of disciplines and related knowledge transfers awaits further work.

REFERENCES

[1] Candle J., Paul D., Turner P., *Business Analysis Techniques,* British Informatics Society (2010)

[2] International Institute of Business Analysis: *A Guide to the Business Analysis Body of Knowledge*, version 3 (2015)

[3] Project Management Institute: *Business Analysis for Practitioners: a Practice Guide* (2015)

[4] Carkenord B.: *Seven Steps to Mastering Business Analysis,* B2T Training (2009)

[5] Brandenburg L., How to Start a Business Analyst Career. www.bridging-the-gap.com (2010)

[6] Madeiros dos Santos P.S, and Travassos G.H., *Action Research Use in Software Engineering: an Initial Survey,* In: Proceedings of the Third International Symposium on Empirical Software Engineering and Measurement, ESEM 2009 (2009), DOI: 10.1145/1671248.1671296

[7] Jacobson I., Ericsson M, Jacobson A.: The *Object Adventure: Business Process Re-Engineering With Object Technology,* Addison Wesley (1994)

[8] JACOBSON I., BOOCH G., RUMBAUGH J.: *THE UNIFIED SOFTWARE DEVELOPMENT PROCESS,* Addison Wesley (1999)

[9] IBM Rational Unified Process Specification, version 7.0.1, 2006, www.ibm.com.

[10] Buckingham M., *StandOut: The Groundbreaking New Strengths Assessment from the Leader of the Strengths Revolution,* Thomas Nelson (2011)

[11] Moczydłowska J.M., Kowalewski K., *Nowe koncepcje zarządzania ludźmi (Novel concepts in HR Management),* Difin (2014)

[12] Ferre X., *Integration of Usability Techniques into the Software Development Process,* In: Proceedings of ICSE 2003 Workshop on Bridging the Gaps Between Software Engineering and Human-Computer Interaction (2003)

[13] Bobkowska A., *Balance Between Creativity and Methodology in Software Projects*, In: Proceedings of the Multimedia, Interaction, Design and Innovation - MIDI '15, ACM DL (2015) DOI:10.1145/2814464.2814468

[14] Boehm B., Turner R., *Balancing Agility and Discipline. A Guide for Perplexed*, Addison-Wesley (2003)

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 265–269
DOI: 10.15439/2016F468
ACSIS, Vol. 9. ISSN 2300-5963

Completeness and Consistency of the System Requirement Specification

Jaroslaw Kuchta
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Narutowicza 11/12, 80-233 Gdansk, Poland
Email: qhta@eti.pg.gda.pl

***Abstract*—Although the System Requirement Specification, as a first formal and detailed document, is the base for the software project in classic software methodologies, there is a noticeable problem of assuring the completeness of this document. The lack of its completeness causes uncertainty of the project foundations. This was one of motivations for agile methodologies – if the SRS cannot be easily validated, if it can change in late project phases, then get rid of the SRS. Replace formal requirements with *user stories*. However *user stories* are also requirements - mostly functional requirements. As agile methodologies focus on functional requirements, it is easy to forget quality requirements.**

In this paper we show the impact of quality requirements analysis on functional requirements exploration. Although in our experiment we noticed considerable large functional requirements increment, we went further and examined the impact of SRS consistency on its completeness. The research has shown that the increment of the revealed requirements count may be almost three times greater, compared to the standard requirement specification method.

Keywords: system requirements, SRS, quality, completeness, consistency,

I. Introduction

Starting the software project, the customers may be uncertain of their expectations, or may simply be unable to imagine the whole complexity of the software system. The developer's task at this very early stage of the project is to reveal as much of the requirements as is possible. In other words – to assure the completeness of the System Requirement Specification. However, there is a noticeable problem with the completeness of the SRS; with its measurement and event with its definition. The formal definition for the requirements completeness says that the SRS is complete when the "*responses of the software to all realizable classes of input data in all realizable classes of situations is included*" [1]. This definition is not very usable, as we do not know the number of "*all realizable classed of input data*" and "*all realizable classes of situations*". Other definitions [2] are also unusable. This leads us to the problem: "*how can we be sure that the requirement specification is complete without knowing what the complete requirement specification is?*" [3]. As it is an impossible situation, we can not determine the absolute completeness; we may only define some metrics of a relative completeness (as the relative increment of the elicited requirements) [4].

The whole system quality depends on the completeness of the requirement specification. We may find a huge set of quality metrics in [5]. As we take a close look at these metrics, we may see that many of them are evaluated relatively to the number of functions described in the requirement specification. As we do not know the completeness of the requirement specification, we cannot be sure about the reliability of these metrics and the total quality of the system.

The uncertainty of the SRS completeness causes the risk of the requirements change during the development process [6]. The risk is extremely high in the classic "waterfall" software development model and it is one of the reasons why the iterative and incremental models became much more popular [7]. Often, an "agile" development process is taken [8], when the requirements are revealed in the user acceptance tests of the working (but incomplete) system. Although this model goes well in the normal situations, there may be some rare and critical conditions, which are hard to reveal based solely on the customer's claims. The author's research has shown, that the problem of the SRS completeness uncertainty may be resolved indirectly – with the measurement the SRS consistency. This measurement (in a graph theory sense) is easy, and has the large impact on the number of the revealed requirements. Using this method, the author has achieved the substantial increment of the requirements (almost three times) compared to the standard method of the requirement specification.

The term "*consistency*" has somehow fuzzy meanings in software engineering: we may understand it as *a lack of contradiction between two, or more, requirements* [9][10]; and also as *traceability*, which means the *ability to trace the requirements to the software solutions* [11][12]. In this paper, we understand consistency as a *degree of coupling* between the requirements. To measure the degree of the internal coupling of the System Requirements Specification, it must be treated not only as a text document, but also in some quasi-graph form; where the vertices represent the requirements, and the edges represent the references between the requirements. The references between the requirements should be created when one requirement *impedes* the other (e.g. the "data backup" requirement impedes the "data restore" requirement). These are called "trace" references. They should also be created when two requirements *complement* each other (e.g. "logging in" complements "logging off"). Other logical references between requirements should be added "manually" during the requirements specification.

There may be several types of requirements: functional requirements, data requirements, and quality requirements. Some of the quality requirements touch the reliability of the system. System reliability depends on the resolution if critical situations which impede the functional requirements. Detection of unresolved critical situations (i.e. not resolved with functional requirements) leads to the assumption that some functional requirements may be missing. However, this assumption must be confirmed (or denied) in the detailed analysis of the SRS.

II. Measuring the SRS Completeness and Consistency

To evaluate the SRS quality we need a set of precise and objective metrics. Traditionally, when an SRS document is text written, we may count some specific weak phrases (as "adequate", "not limited to") [13]. Having the SRS document stored in the quasi-graph form, we defined other metrics of the SRS completeness and consistency.

A. Metrics and Measures

We distinguished metrics and measures in quality measurement. We treat a *metric* as a quality factor to be measured, and a *measure* as a method of the measurement. We divided measures between direct or indirect measures. We counted *direct measures* directly in the document quasi-graph storage; and we calculated *indirect measures* using some formulas. Every direct measure results in some absolute number counted directly by simple graph analysis (e.g. the number of "trace" references). These we called objective measures, as they may be objectively evaluated. Objective measures, although easy to evaluate, are insufficient to express a complete document quality. Some human analysis is needed to reveal the missing requirements, or to find inconsistent ones. The values resulting from this analysis are called expert measures, as the document review must be done by and expert. The review of the SRS should contain a list of missing requirements and a list of quality remarks to already defined requirements. As expert measures are less reliable than objective ones, the usage of expert measures is minimized. Two expert measures are used in this paper: one for completeness and one for consistency. Indirect measure expression usually divides two direct measure results. The result value is scaled from zero to one. Zero means the worst result and One means the best (ideal) result. When the numerator measure gives the number of "negative" elements (e.g. missing elements), then the function for an indirect measure is defined by a formula:

$$F(m,n) = Floor(1 - m/n, 2)$$

where the Floor function rounds the first argument towards zero (with the precision of two decimal digits); m and n are direct measures. Rounding towards zero is needed as the result value of 1.0 can only appear in the ideal situation (e.g. no missing elements).

If one metric is evaluated with several measures, some aggregate function is needed, e.g. a weighted mean function:

$$Avg(m_1, m_2, ...m_n) = \frac{w_1 m_1 + w_2 m_2 + ... + w_n m_n}{\sum_i w_i}$$

where $m_1, m_2, ...m_n$ are component measures, and $w_1, w_2, ...w_n$ are their weights respectively.

Some other terms used in metrics and measures definitions (below) need explanation. These terms are: element, meta-class, relationship, reference and solution. An *element* means an item of the software project (e.g. the system requirement). Each element has its *meta-class*, which describes its possible and required properties and relationships (e.g. FunctionalRequirement is one of the meta-classes). *Relationships* are not only *references*, which are meaningful only in the development process, but also associations meaningful in the runtime. Solution means here an element, or a set of elements, defined with a "trace" reference as the elaboration of some element.

B. SRS Completeness Metrics and Measures

Proposed metrics of the SRS Completeness (CP) are: Formal Completeness (FCP), Semantic Completeness (SCP) and Reference Completeness (RCP) – see fig.1. *Formal Completeness* (FCP) involves *Template Completeness Factor* (TCPF) and *Definition Completeness Factor* (DCPF). First factor (TCPF) tells us, how completely a template of the document is filled. We count the number of elements required by a document template (the number of required meta-classes), and search missing ones. Second factor (DCPF) represents the number of elements with incomplete definition (with one, or more, properties required by meta-classes missing).

Semantic Completeness (SCP) uses an expert measure – *Missing Semantic Element Count* (MSEC). The missing elements must be listed by an expert revising the document. To get the value scaled from Zero to One we must divide MSEC not only by the *Total Semantic Element Count* (TSEC), but by the sum of TSEC and MSEC.

Reference Completeness (RCP) depends on two measures. First; a *solution completeness factor* (SLCF), evaluates the "trace" references leading to the "solutions". Second; an *internal reference factor* (IRFF), evaluates all missing references, that are required by the elements' meta-classes.

C. Consistency Metrics and Measures

Consistency of the SRS (CS) depends on two metrics: Formal Coherence (FCH) and Semantic Consistency (SCS) – see fig. 2.

Formal Coherence (FCH) is measured using graph-theory. When the document (SRS) is represented in a quasi-graph form, nodes represent project elements (requirements) and edges represent relationships (references). Although the references are directed, we evaluate Weak Coherence Factor (WCHF), which is appropriate for undirected graphs (edge direction is examined while measuring correctness – beyond this paper). WCHF is based on Weak Coherence Count (WCHC), which means a count of subgraphs which are

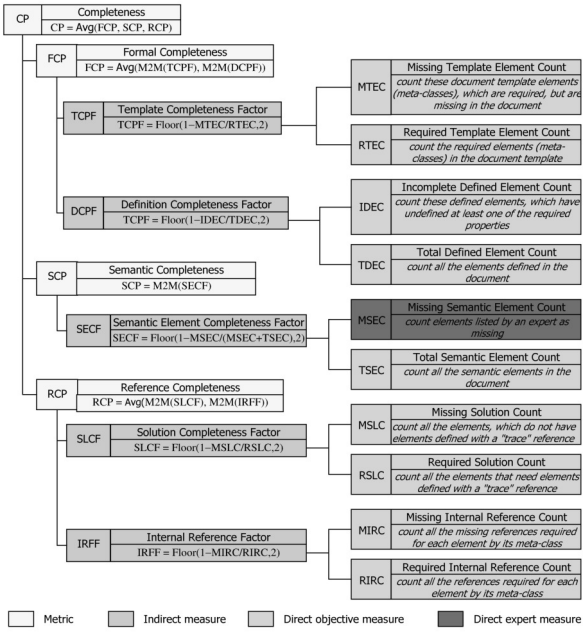


Fig. 1. Completeness metrics and measures

internally coherent, but mutually separate. Besides WCHF, we may also evaluate Relationship Strength Factor (RSTF), which represents the number of bridges in the graph (i.e. references which solely join two parts of the graph).

Semantic consistency (SCS) depends on Semantic Consistency Factor (SCF), which is measured by an expert. The expert must judge if the semantic elements (requirements) are consistent with other elements, and mark inconsistent ones.

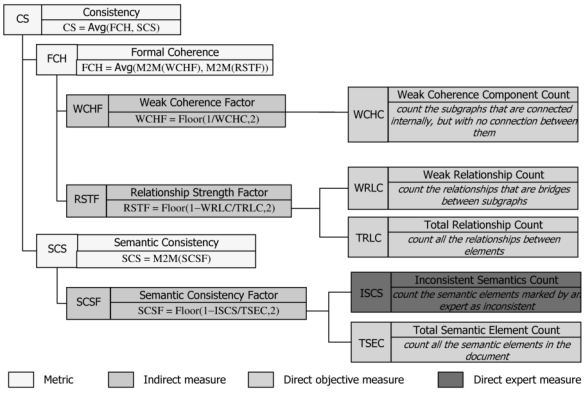


Fig. 2. Consistency metrics and measures

III. EXPERIMENT

A set of metrics and measures for completeness and consistency (shown above) was implemented in the IQuest system, which evaluated not only completeness and consistency, but also correctness, understandability, modifiability and verifiability. However we focus in this paper on the completeness and consistency, as these metrics showed the most interesting coincidence.

The IQuest system allowed us to import a sample SRS document written in Microsoft Office document format as an object-based requirements model where each requirement became an object and these objects were linked with references entered by a requirement manager. We used this object model to evaluate the consistency of a sample requirement specification. The IQuest system evaluated objective quality measures automatically, however an expert had to entered few data to evaluate the expert measures.

A. Evaluating the Quality of a Sample SRS Document

The sample SRS document for a Web Supermarket was elaborated for the research. The specification was presented as the set of requirement specification tables (see fig. 3). Note that references used to evaluate the SRS consistency are marked with '#'

FREQ_19	**Order placing**
Description:	The customer places an order with web service. S/he must find and select an article, determine the quantity and choose a credit card (if has more than one). System determines the total payment amount (price range!) and sends acknowledgement to the customer with the date of deliverance determined.
Applies to:	#USER_01 Customer
Results from:	#USER_01.TASK_04 Ordering
Associations:	#FREQ_15 Article searching #FREQ_17 Article selecting #FREQ_20 Order acknowledging
Sources:	#STKH_01 WDS Directors Board #STKH_07 WDS Sales Director
Priority:	very high

Fig. 3. Example of the functional requirement

Three incrementing versions were prepared and their quality was evaluated. The completeness was decided (for the research) to be the most important factor for requirements specification. The correctness and consistency were decided to be less important and the importance of other metrics was only supplementary. The weights of 40, 20, 20, 10, 5, and 5 were assigned to the quality factors. All the metrics were scaled from 1 (worst) to 6 (best).

The objective measures was not only ones which led to new requirements definition. In the second version also a business expert was asked to find out missing requirements.

B. First version – no quality requirements defined

First version of the SRS document was prepared very thoroughly – as much as 69 functional requirements were specified. Although the quality requirements were not specified, its quality was evaluated. No expert was asked for evaluation because the specification was not finished at that moment. The results are shown in the tab. I.

As no quality requirements were specified, the low value of completeness agreed with expectations.

TABLE I
QUALITY RESULTS OF THE FIRST VERSION OF THE SPECIFICATION

Metric	Weight	Value
Completeness	40	3.1
Consistency	20	5.8
Correctness	20	6
Understandability	10	5
Modifiability	5	5.95
Verifiability	5	5.8
Total quality		4.65

C. The second version – quality requirements consideration

Based on the first SRS, the next version with 36 quality requirements was developed. The quality requirements were revealed according to the template presented by table II. Exceptional, critical and breakdown situations were deliberated for reliability. For each such situation, a functional requirement was specified to prevent, or fix, this situation. That resulted in 99 new functional requirements. The results of the quality evaluation for the second version are shown in table III.

Despite expectation, completeness increased very little (from 3.1 to 3.65). Two explanations were discovered by the detailed result analysis (see table IV). First, it is impossible to achieve high completeness without filling the template completely (FCP). Second, the *Reference Completeness* (RCP) grew, but not satisfying. Despite intensive work, about 30% of elements were left without solution.

D. Third version – consistency consideration

In the third version, the formally "unresolved" goals; advantages, needs, tasks and problems, were deliberated. The emphasis was laid on increasing consistency. Some of the analyzed elements had already solutions in the form of functional requirements, but 30 new requirements had to be specified. The results of quality evaluations are shown in table V.

Although the increment of the consistency was very small (from 5.85 to 5.9), the completeness grew from 3.65 to 5.2. The main reason was the growth of *Reference Completeness* (RCP) from 4.44 to 5.9.

E. Relationship between consistency and completeness

Fig. 4a presents consistency impact on completeness in the three versions of the specification. This impact is significant (not only) in a mathematical aspect. More important is the impact of consistency on the number of functional requirements. As you can see (rys. 4b), this number grew about 3 times (from 69 to 198). It means that without quality evaluation and without consistency increment about 2/3 of functional requirements would be omitted and the requirements specification would be incomplete.

IV. CONCLUSIONS AND FURTHER RESEARCH

Consistency measurement of the Software Requirement Specification requires a quasi-graph representation of this document; with nodes representing requirements, and edges representing references between requirements. The document

TABLE II
THE STRUCTURE OF QUALITY REQUIREMENTS PART OF SRS

8. Quality requirements
8.1. Efficiency requirements
8.1.1. Effectiveness
8.1.2. Load capabilities
8.1.3. Stability
8.1.4. Scalability
8.2. Reliability requirements
8.2.1. Exceptional situations
8.2.2. Critical situations
8.2.3. Breakdown situations
8.2.4. Error resistance
8.2.5. Security and safety
8.2.6. Testability
8.3. Flexibility requirements
8.3.1. Portability
8.3.2. Localizability
8.3.3. Modifiability and configurability
8.4. Usability requirements
8.4.1. Usability
8.4.2. Understandability and learnability
8.4.3. Productivity

TABLE III
QUALITY RESULTS OF THE SECOND VERSION OF SPECIFICATION

Metric	Weight	Value
Completeness	40	3.65
Consistency	20	5.85
Correctness	20	5.9
Understandability	10	5.3
Modifiability	5	6
Verifiability	5	5.8
Total quality		4.9

TABLE IV
DETAIL COMPARISON OF THE COMPLETENESS EVALUATION FOR THE FIRST AND THE SECOND VERSION

Symbol	Metric or measure	v.1	v.2
CP	Completeness	3.1	3.65
FCP	Formal Completeness	4.5	5
TCPF	Template Completeness Factor	0.75	0.87
MTEC	Missing Template Element Count	2	1
RTEC	Required Template Element Count	8	8
DCPF	Definition Completeness Factor	0.94	0.92
IDEC	Incomplete Defined Element Count	19	45
TDEC	Total Defined Element Count	375	578
SCP	Semantic Completeness	-	5.95
SECF	Semantic Element Completeness Factor	-	0.99
MSEC	Missing Element Count	-	3
TSEC	Total Specified Element Count	375	578
RCP	Reference Completeness	4.05	4.44
SLCF	Solution Completeness Factor	0.62	0.7
MSLC	Missing Solution Count	65	71
RSLC	Required Solution Count	176	242
IRFF	Internal Reference Factor	0.99	0.99
MIRC	Missing Internal Reference Count	1	5
RIRC	Required Internal Reference Count	260	577

TABLE V
QUALITY RESULTS OF THE THIRD VERSION OF SPECIFICATION

Metric	**Weight**	Value
Completeness	40	5.2
Consistency	20	5.9
Correctness	20	6
Understandability	10	5.9
Modifiability	5	6
Verifiability	5	5.9
Total quality		5.6

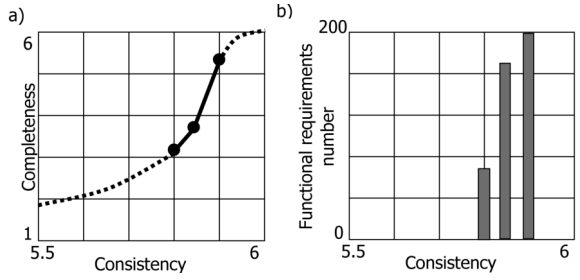


Fig. 4. Consistency impact to completeness (a) and to the number of functional requirements (b) in the three versions of the specification. Dotted line at (a) shows the expected impact.

template defines permitted and required elements that should appear in the document. A set of objective measures may be defined to automate evaluation of the quality of the document. These objective measures are supplemented with expert measures, which are evaluated in the document review process.

The vast impact of consistency on completeness can be noticed while evaluating quality of the SRS document. Consistency does not guarantee completeness, but it may help to reveal much more requirements than using traditional methods.

Basing on this preliminary study we prepared the next experiment. At the first part of the experiment we collected about 50 SRS documents from developers which used the proposed method. Next we gave the same project subjects for other developers using agile methods. We plan to compare the number of functionalities discovered with traditional and with agile methods. However the results will be available after several months.

REFERENCES

[1] "Ieee guide for software requirements specifications," *IEEE Std 830-1984*, pp. 1–26, Feb 1984. doi: 10.1109/IEEESTD.1984.119205

[2] A. Davis, S. Overmyer, K. Jordan, J. Caruso, F. Dandashi, A. Dinh, G. Kincaid, G. Ledeboer, P. Reynolds, P. Sitaram, A. Ta, and M. Theofanos, "Identifying and measuring quality in a software requirements specification," in *Software Metrics Symposium, 1993. Proceedings., First International*, May 1993. doi: 10.1109/METRIC.1993.263792 pp. 141–152.

[3] T. Shell, "System function implementation and behavioral modeling: A systems theoretic approach," *Systems Engineering*, vol. 4, no. 1, pp. 58–75, 2001. doi: 10.1002/1520-6858(2001)4:1<58::AID-SYS6>3.0.CO;2-Z. [Online]. Available: http://dx.doi.org/10.1002/1520-6858(2001)4:1<58::AID-SYS6>3.0.CO;2-Z

[4] R. S. Carson and T. Shell, "Requirements completeness: Absolute or relative? comments on "system function implementation and behavioral modeling"[syst eng 4 (2001), 58-75]," *Systems Engineering*, vol. 4, no. 3, pp. 230–231, 2001. doi: 10.1002/sys.1019. [Online]. Available: http://dx.doi.org/10.1002/sys.1019

[5] S. H. Kan, *Metrics and Models in Software Quality Engineering*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002. ISBN 0201729156

[6] E. Knauss and C. E. Boustani, "Assessing the quality of software requirements specifications," in *2008 16th IEEE International Requirements Engineering Conference*, Sept 2008. doi: 10.1109/RE.2008.29. ISSN 1090-705X pp. 341–342.

[7] R. Pressman, *Software Engineering: A Practitioner's Approach*, 6th ed. New York, NY, USA: McGraw-Hill, Inc., 2005. ISBN 0077227808, 9780077227807

[8] S. Ambler, *Agile Modeling: Effective Practices for eXtreme Programming and the Unified Process*. New York, NY, USA: John Wiley & Sons, Inc., 2002. ISBN 047127190X

[9] "Ieee recommended practice for software requirements specifications," *IEEE Std 830-1998*, pp. 1–40, Oct 1998. doi: 10.1109/IEEESTD.1998.88286

[10] D. Zowghi and V. Gervasi, "On the interplay between consistency, completeness, and correctness in requirements evolution," *the Journal of Information and Software Technology, Volume 45, Issue*, vol. 14, p. 2003, 2003.

[11] T. T. Moores and R. E. M. Champion, "Software quality through the traceability of requirements specifications," in *Software Testing, Reliability and Quality Assurance, 1994. Conference Proceedings., First International Conference on*, Dec 1994. doi: 10.1109/STRQA.1994.526392 pp. 100–104.

[12] G. Kotonya and I. Sommerville, *Requirements Engineering: Processes and Techniques*, 1st ed. Wiley Publishing, 1998. ISBN 0471972088, 9780471972082

[13] W. M. Wilson, L. H. Rosenberg, and L. E. Hyatt, "Automated analysis of requirement specifications," in *Proceedings of the 19th International Conference on Software Engineering*, ser. ICSE '97. New York, NY, USA: ACM, 1997. doi: 10.1145/253228.253258. ISBN 0-89791-914-9 pp. 161–171. [Online]. Available: http://doi.acm.org/10.1145/253228.253258

Software Systems Development & Applications

SSD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution. Events that constitute SSD&A are:

- BTMSPA'16—1st Symposium on Balancing Traditional and Modern Software Process Approaches
- MDASD'16—4th Workshop on Model Driven Approaches in System Development
- MIDI'16 - 4th Conference on Miltimedia, Interaction, Design and Innovation
- SEW-36—The 36th IEEE Software Engineering Workshop

1st Symposium on Balancing Traditional and Modern Software Process Approaches

RESEARCH studies conducted in the area of software engineering have proposed hundreds of techniques and recommendations. On the other hand, software projects are performed in frame of several constraints including required quality, time and budget constraints as well as staff competences. Thus there is a need for selection of the techniques which allow practitioners to achieve a fair balance between time and effort spent on performing these techniques and benefits they provide.

Several traditions exist in software engineering including formal, methodological, user-centered, reuse and agile approaches. Additionally, novel approaches are still being proposed, e.g. collaborative approaches. Each of the approaches focuses on solving a different kind of problems. The proper balance between approaches and, consequently techniques in use, depends on the characteristics of the project at hand.

This symposium aims at exchanging experience and making progress in the knowledge about software process configuration with fair balance between traditional and modern approaches. We invite both experience reports from industry and scientific studies of integrated approaches.

TOPICS

- Balance between agile and disciplined approaches
- Innovation and creativity in software engineering
- Collaborative games in software processes
- Business-IT alignment
- Enterprise integration, business integration and systems integration
- IT-enabled innovations in organizations
- Cooperative, distributed, and collaborative software engineering
- Variability across the software life cycle
- Innovative platforms, architectures and technologies for IS
- Quality assurance and management
- Social media, open data, Internet of Things in business processes
- Methods, tools and human factors in IS/IT management
- Industrial case studies and experience reports related to the above topics

EVENT CHAIRS

- **Bobkowska, Anna,** Gdansk University of Technology, Poland
- **Brudło, Piotr,** Gdansk University of Technology, Poland
- **Przybyłek, Adam,** Gdansk University of Technology, Poland

PROGRAM COMMITTEE

- **Alshayeb, Mohammad,** King Fahd University of Petroleum and Minerals, Saudi Arabia
- **Bauer, Veronika,** Technische Universität München, Germany
- **Belle, Alvine Boaye,** École de Technologie Supérieure, Canada
- **Blech, Jan Olaf,** RMIT University, Australia
- **Borg, Markus,** SICS Swedish ICT AB, Sweden
- **Chatzigeorgiou, Alexandros,** University of Macedonia, Greece
- **Czarnecki, Krzysztof,** Gdańsk University of Technology, Poland
- **Diebold, Philipp,** Fraunhofer IESE, Germany
- **Gregory, Peggy,** University of Central Lancashire, United Kingdom
- **Jasek, Roman,** Tomas Bata University in Zlin, Czech Republic
- **Kaloyanova, Kalinka,** Sofia University, Bulgaria
- **Kapitsaki, Georgia,** University of Cyprus, Cyprus
- **Katić, Marija,** School of Computing, Engineering and Physical Sciences, United Kingdom
- **Knodel, Jens,** Fraunhofer IESE, Germany
- **Kuchta, Jarosław,** Gdansk University of Technology, Poland
- **Madeyski, Lech,** Wrocław University of Technology, Poland
- **Mangalaraj, George,** Western Illinois University, United States
- **Merunka, Vojtech,** Czech Technical University in Prague (Associated Professor), Czech Republic
- **Molhanec, Martin,** Czech Technical University in Prague, Czech Republic
- **Morales Trujillo, Miguel Ehecatl,** National Autonomous University of Mexico, Mexico
- **Nawrocki, Jerzy,** Poznan University of Technology, Poland
- **Norta, Alex,** Tallinn University of Technology, Estonia
- **Noyer, Arne,** University of Osnabrueck and Willert Software Tools GmbH, Germany
- **Özkan, Necmettin,** Turkiye Finans Participation Bank, Turkey
- **Pereira, Rui Humberto R.,** Instituto Politecnico do Porto, Portugal
- **Przechlewski, Tomasz,** Powiślańska Szkoła Wyższa w Kwidzynie, Poland

- **Ramsin, Raman,** Sharif University of Technology, Iran
- **Salnitri, Mattia,** University of Trento, Italy
- **Santos Neto, Pedro de Alcântara dos,** Universidade Federal do Piauí, Brazil
- **Śmiałek, Michał,** Politechnika Warszawska, Poland
- **Soares, Michel,** Federal University of Sergipe, Brazil
- **Soja, Piotr,** Cracow University of Economics, Poland
- **Tarhan, Ayca,** Hacettepe University Computer Engineering Department, Turkey
- **Wiszniewski, Bogdan,** Gdansk University of Technology, Poland
- **Zarour, Nacer Eddine,** University Constantine2, Algeria
- **Zeid, Amir,** American University of Kuwait, Kuwait

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 275–282
DOI: 10.15439/2016F587
ACSIS, Vol. 9. ISSN 2300-5963

Using LINQ as a universal tool for defining architectural assertions

Bartosz Frąckowiak
Institute of Informatics
University of Warsaw
Poland
Email: b.frackowiak@mimuw.edu.pl

Robert Dąbrowski
Institute of Informatics
University of Warsaw
Poland
Email: r.dabrowski@mimuw.edu.pl

***Abstract*—We demonstrate that Microsoft LINQ can be used as a convenient tool to define architectural assertions. We introduce an abstract model of software based on a directed multi-graph and formalize the notion of software architecture and architectural assertions. We demonstrate how Microsoft Visual Studio can be harnessed to extract the architecture of a given software project and append it with assertions using LINQ notation. In particular we explain the flow of data processing that takes place within Visual Studio engine. We follow with examples of assertions selected to demonstrate the expressive power of our approach. We conclude by showing subsequent areas of research worth following in order to deepen the research indicated in this paper.**

I. Introduction

SOFTWARE engineering is concerned with development and maintenance of software systems. Properly engineered systems are reliable, satisfy user requirements while their development and maintenance is affordable.

In the past half-century computer scientists and software engineers have come up with numerous ideas for how to improve the discipline of software engineering. Edgser Dijkstra in his article [11] introduced structural programming which restricted imperative control flow to hierarchical structures instead of *ad-hoc* jumps. Computer programs written in this style were more readable, easier to understand and reason about. Another improvement was the introduction of the object-oriented paradigm [19] as a formal programming concept. Other improvements in software engineering included e.g. engineering pipelines and software testing.

In the early days software engineers perceived significant similarities between software and civil engineering processes. The waterfall model [23] that resembles engineering practices was widely adopted as such, even though despite its original description actually suggesting a more agile approach. It has soon turned out that building software differs from building skyscrapers and bridges. In the late 1990s the idea of extreme programming emerged [3], its key points being: keep the code simple, review it frequently and test early and often. Among numerous techniques, test-driven development was promoted, which eventually resulted in increased quality of produced software and the stability of the development process [14]. Contemporary development teams started to lean towards short iterations (sprints) rather than fragile upfront designs, and short feedback loops allowed customers' opinions to provide timely influence on software development. This allowed for creating even more complex software. Growing complexity of software required ability to describe the software on different levels of abstraction, and the notion of software architecture has developed. The emergence of patterns and frameworks had a similar influence on architecture as design patterns and idioms had on programming. Software became developed by assembling reusable software components, that interact using well-defined interfaces, while component-oriented frameworks and models provided tools and languages making them suitable for formal architecture design. However, a discrepancy between architecture level of abstraction and programming level of abstraction prevailed. While the programming phase remained focused on generating code within a preselected (typically object-oriented) programming language, the architecture phase took place in the disconnected component world. The discrepancies deepened as software gained features while not being properly refactored, development teams changed over time, worked under time pressure with incomplete documentation and requirements that were subject to frequent changes. Multiple development technologies, programming languages and coding standards made this situation even more severe. Unification of modeling languages failed to become the silver bullet.

The discrepancy accelerated research on software architectures, model-driven development or automated software engineering. Nowadays, we have ideas of how to craft the architecture, though we still require ways to both monitor the state of the architecture and enforce it during programming in an automated manner. This is the problem that we aim at in our reseach.

We start with a new vision for management of software architecture based on the idea of an architecture warehouse. *An architecture warehouse* is a repository of all software system and software process artifacts. Such a repository can capture architecture information which was previously only stored in design documents or simply in the minds of developers. *Software intelligence* is a tool-set for analysis and visualization of this repository's content [7], [8], [9]. That includes all tools able to extract useful information from the source code and other available artifacts (like version control history).

All software system artifacts and all software engineering process artifacts being created during a software project are represented in the repository as vertices of a *graph*. Multiple edges of this graph represent various kinds of dependencies among those artifacts. The key aspects of software production like quality, predictability, automation and metrics are then expressed in a unified way using graph-based terms.

The integration of source code artifacts and software process artifacts in a single model opens new possibilities. They include defining new metrics and qualities that take into account all architectural knowledge, not only the knowledge about source code. The state of software (the artifacts and their metrics) can be conveniently visualized on any level of abstraction required by software architects (i.e. functional level, package level) or by software programmers (i.e. class or method level).

Furthermore, the relations among those artifacts can be automatically governed, in particular by implementing the idea of assertions at the architectural level of abstraction.

In this article we introduce concepts and tools that allow architects to enforce architectural principles (constraints) upon programmers using architectural assertions, and we demonstrate their proof-of-concept implementation using Microsoft LINQ and Microsoft Visual Studio.

We introduce a new way of using internal Visual Studio components to provide a universal tool for discovering violation of architecture constraints. Typically, tools of this type provide functionality via a new standalone platform, doubling existing functionality of integrated development environments; or at best get integrated with existing environments (i.e. as their plugins). We take a different approach. Since developers spend most of their time using integrated development environments as the main tool for producing source code, we aim at reusing as much functionality of the developers' well known environment as possible. In this approach LINQ becomes a universal language for describing architectural assertions for all types of programming languages, and Visual Studio becomes a universal environment with software intelligence capabilities extending beyond its natively supported programming stack.

The paper is organized as follows. Section II briefly summarizes the works related to this research. Section III recalls the graph-based model for representing architectural knowledge that creates the backbone for architectural assertions, while Section IV describes their implementation using Visual Studio and LINQ. Section V demonstrates by example how this approach can be applied to handle selected architectural challenges. Section VI concludes.

II. Related work

In 2010 Tibermacine et al. [29] worked on a family of languages for architecture constraint specification.

They argued that during software development architectural decisions should be documented so that quality attributes guaranteed by these decisions and required in the software specification could be preserved. They stressed out that an important part of these architectural decisions is getting them formalized using constraint languages which differ between stages of the development process. Therefore they suggested a family of architectural constraint languages, where each member of the family, called a profile, was to be used to formalize architectural decisions at a given stage of the development process. All profiles were based on a certain core constraint language and a common architecture model. In addition to the family of languages, they introduced a transformation-based interpretation method for profiles and an associated tool.

In 2012 Fabresse et al. [13] have worked on bridging the gap between design and implementation. They observed that significant amount of software systems are designed in component-oriented approach but programmed in object-oriented languages. Unified Modeling Language (UML), Corba Component Model (CCM) or Enterprise Java Beans (EJB) were shown as examples of component-oriented models that were only used at design time, while implementation relied on object-oriented languages, with developers not actually adapting component-oriented programming. The authors identified decoupling, adaptability, unplanned connections, encapsulation and uniformity as important requirements for component-oriented programming and proposed a language that fulfilled these requirements, along with a prototype implementation and concrete experiments to validate their proposal.

As software evolution has become an integral part of the software lifecycle, Lytra et al. [16] focused their research on checking consistency between design decisions and design models. In 2012 they proposed a constraint-based approach for checking the consistency between the decisions and the corresponding component models. They argued that since maintenance of a software system involves among others the maintenance of the software system architecture, then software community must come up with additional models to capture architectural design decisions and their design rationale, and record the architectural knowledge. Their approach enabled explicit formalized mappings of architectural design decisions onto component models. Based on these mappings, component models along with the constraints used for consistency checking between the decisions and the component models were to be automatically generated using model-driven techniques. The approach was coping with changes in the decision model by regenerating the constraints for the component model. Thus, the component model got updated and validated as the architectural decisions evolved.

In 2013 and 2014 Spacek et al. [24], [25], [26] worked on wringing out objects for programming and for modeling of component-based systems, and bridging the gap between component-based design and implementation with a reflective programming language. They recalled that languages and technologies used to implement component-based software are not component-based, i.e. while the design phase happens in the component world, the programming phase occurs in the object-oriented world; and when an object-oriented language is used for the programming stage, then the original component-based design vanishes, because component concepts are not

treated explicitly. They suggested a pure reflective, component-based programming and modeling language, where all core component concepts were treated explicitly and therefore kept the original component-based design alive. The language made it possible to model and program software using the same language, while its uniform component-based meta-model and integrated reflection capabilities aimed at making the language and its applications flexible.

III. Model

In our work we extend research summarized in sections I and II.

We follow the unified representation of software architecture as a collection of artifacts created during a software (development) process and the relations among those artifacts. We model it with a directed labeled multigraph [8].

Such model caters for the following key needs: (1) natural scalability, (2) abstraction from programming paradigms, languages, specification standards, testing approaches, etc, and (3) completeness, i.e. all software system and software process artifacts are represented.

Our goal is to ensure that the designed architecture is kept on track during the whole software process, in particular can be enforced upon programmers during software development.

We obtain this goal by harnessing Microsoft technology to deliver tools that allow to: (1) define architectural assertions in concise notation; (2) monitor breaking the assertions by programmers in automated way.

A. Architecture graph

Let *software architecture* be the structure or set of structures defined by existing software elements, and the relationships among them, best represented by *software architecure graph*.

Let *software architecture graph* be an ordered tuple:

$$(\mathcal{V}, \mathcal{E})$$

where $\mathcal{V}$ is the set of vertices that reflect design and implementation artifacts created during a software project; $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of directed edges that represent dependencies (relationships) among those artifacts.

For simplicity of reasoning, in this paper we limit the model in respect to its original definition [7] by restricting the range of information about a project being collected by the architecture graph.

Let vertices of our architecture graph be limited to the following types:

$$\mathcal{V} = \{module; class; method\}.$$

These are the artifacts that are typically available in most modern object-oriented languages (including lambda expressions being anonymous methods).

Let edges of our architecture graph be limited to the following types:

$$\mathcal{E} = \{association : class \rightarrow class;$$
$$association : class \rightarrow method;$$
$$association : method \rightarrow class;$$
$$association : method \rightarrow method;$$
$$creation : method \rightarrow class;$$
$$inheritance : class \rightarrow class;$$
$$call : method \rightarrow method\}.$$

For example: an association denotes that a method is contained within (owned by) a class; also association relation exists when a method takes or returns as an argument an instance of a different class; or when a class defines a property within a different class; creation represents special methods responsible for construction statements in the source code, ie. using *new* keyword; inheritance denotes class hierarchical relations; call denotes steering being transferred from one method to another.

For the simplicity of the reasoning, in this paper we assume only static relations are represented in the architecture graph; though dynamic aspects, ie. dynamic calls, are possible to be automatically discovered and represented in the model [17].

We also omit types of static relations among artifacts, ie.

$$inclusion : module \rightarrow module \notin \mathcal{E}.$$

However please note that our approach to implementation of architectural assertions makes extending the scope of architectural knowledge represented in our architecture graph's easy; for more details on possible extensions see section VI. Also please note, that in our approach the relations can be implicitly extended by folding existing relations into new types of relations, ie.:

$$association : class \rightarrow method$$

and

$$creation : method \rightarrow class$$

in fact define

$$classcreation : class \rightarrow class,$$

that is a creator relation in which one class is responsible for creating objects of another class; see example of class factory in section V.

B. Architecture assertion

Let source project $\mathcal{P}$ be a software project created in any modern programming language (typically object-oriented) that is to be constrained using architectural assertions. Let $\mathcal{G} = \mathcal{G}(\mathcal{P})$ be the architecture graph derived for the project $\mathcal{P}$ (extracted from the project's source code).

In the remaining part of the paper please observe, that though we implement our approach using Visual Studio tools, this does not restrict the range of languages that our approach can be applied to.

Let architecture query denote a function that returns a subgraph of the given architecure graph

$$\mathcal{Q} : \mathcal{G} \rightarrow \mathcal{G}'$$

where $\mathcal{G}$ and $\mathcal{G}'$ are architecture graphs and $\mathcal{G}' \subseteq \mathcal{G}$. Then we can define architecture assertion as a comparison of the result set of an architecture query to the empty set.

Let architectural assertion $\mathcal{A}$ denote such an architecture query that the assertion is met (true) iff the executed query returns an empty graph; otherwise the assertions is broken (false):

$$\mathcal{A} : \mathcal{Q} \rightarrow \{true, false\}$$

defined as

$$\mathcal{A}(\mathcal{Q}) := \mathcal{Q}() == \emptyset \ ? \ true : false.$$

C. Architecture processing

We assume that tasks of software architects include defining constraints that bind software programmers during software development process. Put otherwise, architects create assertions that define desired (and also undesired) relations between the components of the system. A library of such assertions, when created, contributes to project's architectural knowledge. Consequently, the general approach to processing architectural assertions is as follows.

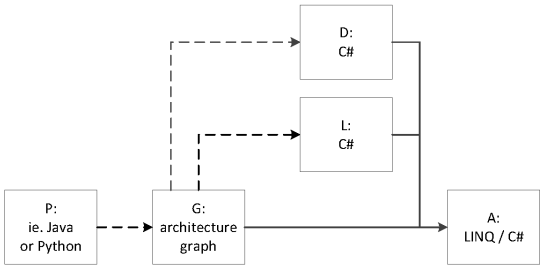


Figure 1. General approach to assertion processing

1) Process the source project $\mathcal{P}$ to extract its architecture graph $\mathcal{G}$;
2) Process the architecture graph $\mathcal{G}$ into the design project $\mathcal{D}$ denoted in a domain-specific language $\mathcal{L}$;
3) Create a collection of architectural assertions $\mathcal{A}$, where each assertion is defined in $\mathcal{L}$ in a notation referring to design project $\mathcal{D}$ (being an abstraction of the source project $\mathcal{P}$);
4) Evaluate assertions $\mathcal{A}$ to identify in $\mathcal{D}$ breaches of architectural rules;
5) Retract from $\mathcal{D}$ via $\mathcal{G}$ into $\mathcal{P}$ to identify fragments of $\mathcal{P}$ that violate architectural constraints imposed on the project.

See the following section IV for details on how those concepts have been assembled together using Microsoft technologies to constitute a general-purpose tool for defining and monitoring architectural assertions; see section V for examples of assertions.

IV. Implementation

Please recall the key design concept introduced in section III: (1) extracting from a given source project an abstract model that focuses only on architectural artifacts and their relations; (2) expressing the artifacts and relations in an intermediary layer denoted in a domain-specific language; (3) using an existing calculation environment capable of processing given domain-specific language as its calculation input.

For our proof-of-concept implementation of the design concept described in section III we harness the following Microsoft components:

- **IDE** Visual Studio Integrated Development Environment providing graphical user interface framework we extend for our purposes;
- **DSL** Visual Studio Domain Specific Language Tools allowing us to define an own domain-specific language to represent an abstraction of the source code;
- **T4** tools providing processing and persistence capabilities for our abstraction of the source code;
- **LINQ** syntactic sugar for concise notation of architectural assertions, ie. thanks to using anonymous methods (lambda queries);
- **Roslyn** for on-the-fly parsing, compiling and executing of the assertions.

A high-level overview of processing steps is depicted on figure IV. In subsequent parts of the section we provide more details on the goals of each step and how the components we selected are used to achieve those goals. We stress out that using these components, especially LINQ as the assertion notation, proves to be efficient in terms of: (1) high expressive power of notation used to define architectural assertions; and (2) small programming effort required to implement the automated verification of such assertions.

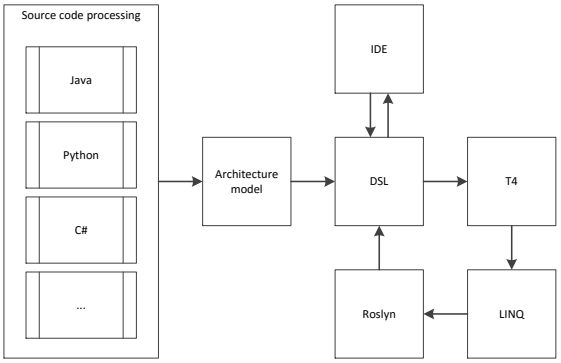


Figure 2. Implementation of processing steps using Microsoft components

A. Source code processing

A prerequisite for further adding and executing architectural assertions is processing source code (of the source project) in order to obtain its architectural model. In many cases such processing can be implemented effectively by analyzing the source code's abstract syntax trees (AST); ie. in case of languages like Java, Pythor or $C\#$ the compilers (interpreters) allow to analyze the source code's ASTs.

When examining a single node (of the tree) representing a method, we deduce the fact that the method calls another

method (in some other class). Next we perform type solving; through a preliminary compilation of sources we check what is the type (class) of this method. Please note that though pre-compilation process is great for collecting information about software archiecture, there might be some cases when pre-compilation is not possible (ie. in case of interpreted programming languages). In this research we narrow focus only to languages for which source code compilers of adequate capabilities exist, and assume in our approach that abstract syntax trees constitute a first layer of abstraction between the source code files and the architecture graph.

B. DSL

To represent source code abstraction collected during source code processing, we utilize Visual Studio and its ability to provide an abstract mechanism for representing structures of any selected domain; namely we implement our architecture model using its Domain Specific Language Tools. To implement the graph structure we extend the DSL Tools' interfaces with own implementation classes, main ones being as follows.

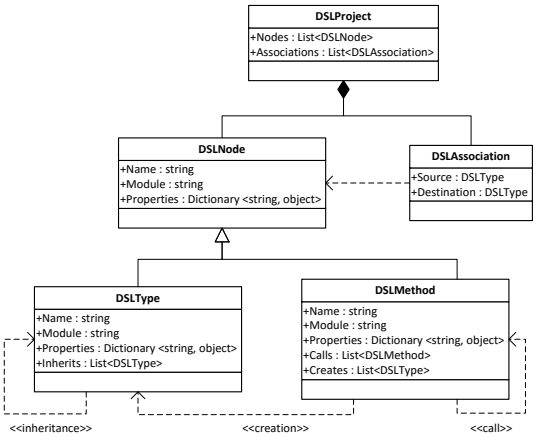


Figure 3. DSL representation of architecture graph

1) **DSLProject** is a container for all the artifacts of architectural model and the relations among them.
2) **DSLNode** is the superclass for all types of artifacts, sharing common properties of further entities. Those common properties are: (1) name (string), (2) module (string), and (3) dynamic dictionary of artifact's properties.
3) **DSLAssociation** allows to represent all type of ownership relations among the artifacts, like classes owning methods, methods owning (anonymous) methods, methods owning (anonymous) classes, classes owning (anonymous) classes.
4) **DSLType** represents the object-oriented concept of class entity, or any other similar architectural artifact, like structure, enumeration, etc.
5) **DSLMethod** represents architectural artifact for method statements in the software source code. Please note that to represent relation of binding methods to a class (like in object-oriented source code) we use DSLAssociation; however for several special relations (like creation, inheritance or call) we made a design decision of keeping direct pointers.

C. T4

In the next processing step we transform the DSL-represented data using T4 templates. The templates use instances of objects described in a DSL model as the input and generate any text containing the data from the DSL model. Templates allow iterations through data collections, conditional statements and text transformation using any software libraries provided for the .NET platform. In our case the templates are used to generate an intermediary code (in $C\#$) to be next appended with architectural assertions (in LINQ) and interpreted (with Roslyn). Regardless of the programming language of the source project, there are three types of $C\#$ files that get generated (for one source project).

1) **Architecture Model** The first file contains definition of artifacts that are used in definitions of architectural queries. Put otherwise, it contains the data model derived from the source code that gets referred to by architectural assertions. Technically, its is a reflection of the DSL representation of architecture graph. By design it is constructed to provide read-only data, ie. it should not be directly edited by software architects; if needed it should be re-generated from the source code. In some sense it resembles definitions of the objects in DSL layer with methods omitted (ie. status changing methods).
2) **Intermediary Code** While the first file contains an abstraction of the source code, the second file contains $C\#$ translation of the given source project. In particular the $C\#$ code contains the same structures as the corresponding structures in the source project, ie. if the source code includes some class, then in the intermediary $C\#$ code a class with exactly same name and properties will be created. The code inside the second file refers to the definitions from the first file. The second file is required for eventual execution of queries againts the source code that the queries aim to constraint. This file is also generated automatically from the source code and should not be edited manually by software architects.
3) **Query Definitions** The third file contains a collection of architectural queries. Definition of each query refers to definitions from the first file. Queries can be freely edited by architects, according to their personal experience. For the definitions of the queries to be interpreted correctly by VS, it must be combined with the previous two files.

Please note, that the first two files are delivered as a result of analysis of the source code and its transformation into $C\#$. They provide resources for software architects to define next their own architectural assertions. The resources are provided read-only, that is in case of architectural changes in the source code, the intermediary $C\#$ files must be re-generated. On the other hand, the third file contains a collection of queries as defined by architects themselves. Software architects are

encouraged to build a collection of re-usable queries, so that introducing architectural rules or restrictions into new projects becomes quick and unexpensive.

D. Roslyn

Eventually the assertions are executed using Roslyn. It begins with combining all three types of files created for the given source project as described in the previous subsection into a $C\#$ program, and reparsing the respective $C\#$ program. What follows is an on-the-fly compilation of the program; compilation on-the-fly is a fully-fledged compilation, the same as for the creation of library files or other executables, with the exception that the compilation unit immediately goes into memory and is managed with a current thread. Within execution of the current thread it is possible to call any method of the loaded library and retrieve results, though architecture query results must be of types available in the unit of compilation.

In solution proposed in this paper, an architecture query result (graph) can be represented using simple types. That is the query may return either a tuple being a subset of vertices (list of strings) and a subset of edges (list of strings) of the architecture graph; or the subset of edges is empty, with a nonempty subset of vertices; or both subsets are empty (empty graph, with both lists being empty).

V. Main result

Please recall that an assertion is a comparison of an architecture query result to an empty set, where architecture query is a function returning a subset of the project's architecture graph. Hence the assertion is satisfied if and only if the query returns an empty graph. We demonstrate by the following examples how our approach can be applied to enforce architectural assertions:

(1) Unwanted Instability: only stable or unstable classes are allowed; (2) Factory Violation: object are constructed in factories only; (3) God Object: all-powerful objects are not allowed. For each example we summarize: (1) the architectural *problem* it aims to solve; (2) the subset of the graph *model* relevant for the assertion; and (3) the definition of the resulting *query*.

Please note the concise notation used to denote the query. Also please recall, that using this approach the assertions can be denoted in LINQ and $C\#$ for any type of source code, as long as the source code can be abstracted into a unified model of architecture graph; in particular it suffices that for the given source code language there exists an AST toolset, as it does ie. in respect to Java or Python.

A. Unwanted Instability

Problem: Instability metric (I) indicates module, package or class readiness for change [18]. It is calculated as the ratio of efferent coupling (Ce) to the efferent and afferent coupling (Ca), namely $I = Ce/(Ce+Ca)$. The range for this metric is $I \in [0..1]$. Accorind to metric author a module with instability close to value 0 is considered stable. A stable module has no references to other modules, can have number of internal references (among module's own artifacts). On the other hand a module with instability metric value close to 1 is considered instable. An instable module usually has a vast amount of outgoing references and a low amount of internal references. Modules with instability $I \in (0.3..0.7)$ are considered neither stable or unstable, as such being typically unwanted in a software project. The following query finds out which modules are unwanted in terms of instability metric (as defined above).

Model:

$$\mathcal{V} = \{class; method\}.$$

$$\mathcal{E} = \{association : class \rightarrow method;$$

$$call : method \rightarrow method\}.$$

Query:

```
[AssertAttribute]
public static IEnumerable<string>
      UnwantedInstability(){
  var en = from types in Project.Types.Where(x =>
    x.Callers.Count + x.Calls.Count > 0
  )
  group types by new {
    types.FullName,
    Ca = types.Callers.Count,
    Ce = types.Calls.Count,
    I = types.Calls.Count /
       (double)(types.Callers.Count +
       types.Calls.Count)
  }
  into rType
  where
    rType.Key.I >= 0.3 &&
    rType.Key.I <= 0.7
  select rType.Key.FullName;

  return en.ToList();
}
```

B. Factory Violation

Problem: The purpose of factory pattern is to hide the logic of creating individual objects. Creating objects outside the designated class factories violates the pattern. We assume that in our architecture model the factories are explicitly indicated, that is each such artifact has a name that contains *factory* suffix. We search for violations of factory pattern.

Model:

$$\mathcal{V} = \{class; method\}$$

$$\mathcal{E} = \{association : class \rightarrow method;$$

$$creation : method \rightarrow class;$$

$$call : method \rightarrow method\}$$

Query:

```
[AssertAttribute]
public static IEnumerable<string> FactoryViolation() {
  const string factoryLabel = "factory";

  var v = Project.Types
    .Where(x => x.Name.ToLower()
    .Contains(factoryLabel));

  if (!v.Any())
    return new[] { "" };

  var factoryArtifacts = Project.Types
    .Where(
      conn => conn is CreationConnection &&
      v.Any(y => y == x.Source)
    ).Select(x => x.Destination);

  return Project.Types
    .Where(
      conn => conn is CreationConnection &&
      factoryArtifacts.Any(y => y == x.Destination) &&
      !x.Source.Name.Contains(factoryLabel)
    );
}
```

C. God Object

Problem: One of previous examples shows how to use our approach to find out a violation of design patterns. Same idea can be used to find out if anti-patterns exist in the source project. One of well-known anti-patterns is *God Object*. This anti-pattern is a violation of *Single responsibility principle* rule defined as a part of SOLID rules [18]. Single responsibility principle constrains one class to provide logic for only just one functionality. Keeping classes inside source code simple and responsible for one thing each increases source code maintainability and software scalability. Placing one class which does too much things is a tempting phenomenon for inexperienced developers, hence we search for existence of *God Objects*.

Model:

$$\mathcal{V} = \{class; method\}$$

$$\mathcal{E} = \{association : class \rightarrow class\}$$

Query:

```
[AssertAttribute]
public static IEnumerable<string> GodObject()
{
  var types = Project.Types.Select(x => new
  {
    Type = x,
    Count = Project.Connections.
      Count(y => y.Destination == x || y.Source == x)
  });
  var v = types.OrderByDescending(x => x.Count);
  return v.Take(3).Select(x => x.Type.FullName);
}
```

VI. Conclusions

Our paper follows the research on architecture of software and software process. It promotes an approach that avoids separation between source code, software process and software architecture (design) artifacts. In this paper we demonstrate that an implementation of such approach is feasible. We demonstrate that LINQ, being a syntactic extension of $C\#$, can become a concise and expressive notation for defining architectural assertions. We also demonstrate that Visual Studio, being an integrated development environment, is a good platform to create a tool that allows software architects to enforce assertions upon software projects, uniformly treating source and architectural layers of the project.

The idea to extend functionality of existing integrated development environments is not novel, it has been alredy confirmed in practice and there exist plugins for specific domains of software engineering. The actual novelty of our approach lies in representing architectural artifacts using the same artifacts as the ones in the source code. More precisely, in parallel to the source project we generate an additional design project describing architecture of the source project. Additinally, if both projects follow the same syntactical rules (of the same programming language, ie. $C\#$ like in our example), then architectural artifacts can even melt with the actual source code hence be accessed and automatically processed just as any other parts of the source code, both from the perspective of integrated development environment, and from the perspective of a software programmer or software architect. Parallel execution of both projects - the source project and the design project - opens new opportunities. While the first project preserves its original business purpose, the second project becomes responsible for watching over internal architecture of the first project - validation of the architectural constraints placed upon artifacts and their interconnections. Another novel observation is the fact that to denote and automatically validate architectural assertions, software architects do not need to explore all the details of the source code. For this purpose only a certain abstraction of the source code is satisfactory - focusing on signatures of types and methods, their relations, but disregarding implementation specifics, ie. method conditional statements. Therefore we have observed that a source project in programming language A (ie. Java or Python, or any other language of similar characteristics) can be automatically transformed into a design project in another programming language B (ie. in $C\#$) such that we can express in language B architectural assertions in respect to architectural artifacts of the project in language A. It is actually practicable due to a common abstraction behind majority of current object-oriented programming languages. Thanks to this, for ie. a program in Java we can denote architectural assertions in LINQ, melt them into an automatically generated $C\#$ abstraction of the Java program and compile and execute the new $C\#$ program to get the assertions validated. In some sense this way $C\#$ becomes a *calculation description* and Visual Studio a *calculation engine*.

Our research can be extended in a few directions. Practical

aspects include creating a publicly-available extension for Visual Studio containing all the described functionality, integrated and operationally verified. Theoretical aspects include extending the scope of graph model of software architecture (available as the domain for the architecture assertions) to include subsequent artifacts and subsequent relations; also researching its properties, expressive power, and the scope of programming languages compatible with this model; generally expanding the concept of collecting architectural knowledge [8], [1], [28]. Our approach requires also thorough verification and comprehensive, comparative testing; in particular creating a publicly available test-bed consisting of multiple source projects in multiple programming languages is a must. It can be anticipated that examples that include reflections, functional programming, dynamic method definition or other programming concepts may require refactoring of DSL implementation of the model, or redefinition of T4 transformations used for model processing, or introducing other concepts of model transformations [10]. In parallel we intend to start building a default library of architectural assertions denoted in LINQ, that would cover existing good architectural and design practices [4], [5]. Such library - created in abstraction from the source code language - would trigger another research stream, namely existing software projects could be systematically verified against the predefined architectural assertions. Another topic of research is appending information of architectural assertions into visual representation of software as a graph [2], [6], [15], [20]. Yet another direction is empowering software architects with tools (hence a notation) that would implement the long defined postulate that software process is a software as well [21], or even extended such approach to actually include all software project artifacts, not only source code artifacts [22], [27], [12].

REFERENCES

[1] M. A. Babar and I. Gorton, editors. *Software Architecture, 4th European Conference, ECSA 2010, Copenhagen, Denmark, August 23-26, 2010. Proceedings*, volume 6285 of *Lecture Notes in Computer Science*. Springer, 2010.

[2] C. Bartoszuk, R. Dąbrowski, K. Stencel, and G. Timoszuk. On quick comprehension and assessment of software. In B. Rachev and A. Smrikarov, editors, *CompSysTech*, pages 161–168. ACM, 2013.

[3] K. Beck. Embracing change with extreme programming. *IEEE Computer*, 32(10):70–77, 1999.

[4] H. P. Breivold, I. Crnkovic, and M. Larsson. Software architecture evolution through evolvability analysis. *Journal of Systems and Software*, 85(11):2574–2592, 2012.

[5] N. Brown, R. L. Nord, I. Ozkaya, and M. Pais. Analysis and management of architectural dependencies in iterative release planning. In *WICSA*, pages 103–112, 2011.

[6] C. S. Collberg, S. G. Kobourov, J. Nagra, J. Pitts, and K. Wampler. A system for graph-based visualization of the evolution of software. In S. Diehl, J. T. Stasko, and S. N. Spencer, editors, *SOFTVIS*, pages 77–86, 212–213. ACM, 2003.

[7] R. Dąbrowski. On architecture warehouses and software intelligence. In T.-H. Kim, Y.-H. Lee, and W.-C. Fang, editors, *FGIT*, volume 7709 of *Lecture Notes in Computer Science*, pages 251–262. Springer, 2012.

[8] R. Dąbrowski, K. Stencel, and G. Timoszuk. Software is a directed multigraph. In I. Crnkovic, V. Gruhn, and M. Book, editors, *ECSA*, volume 6903 of *Lecture Notes in Computer Science*, pages 360–369. Springer, 2011.

[9] R. Dąbrowski, G. Timoszuk, and K. Stencel. One graph to rule them all software measurement and management. *Fundam. Inform.*, 128(1-2):47–63, 2013.

[10] J. Derrick and H. Wehrheim. Model transformations across views. *Sci. Comput. Program.*, 75(3):192–210, 2010.

[11] E. W. Dijkstra. Letters to the editor: go to statement considered harmful. *Commun. ACM*, 11(3):147–148, 1968.

[12] A. Egyed and P. Grünbacher. Automating requirements traceability: Beyond the record & replay paradigm. In *ASE*, pages 163–171. IEEE Computer Society, 2002.

[13] L. Fabresse, N. Bouraqadi, C. Dony, and M. Huchard. A language to bridge the gap between component-based design and implementation. *Computer Languages, Systems & Structures*, 38(1):29–43, 2012.

[14] R. Kaufmann and D. Janzen. Implications of test-driven development: a pilot study. In *Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, OOPSLA '03, pages 298–299, New York, NY, USA, 2003. ACM.

[15] R. Koschke. Software visualization for reverse engineering. In S. Diehl, editor, *Software Visualization*, volume 2269 of *Lecture Notes in Computer Science*, pages 138–150. Springer, 2001.

[16] I. Lytra, H. Tran, and U. Zdun. Constraint-based consistency checking between design decisions and component models for supporting software architecture evolution. *2011 15th European Conference on Software Maintenance and Reengineering*, 0:287–296, 2012.

[17] V. Markovets, R. Dąbrowski, G. Timoszuk, and K. Stencel. Know thy source code. is it mostly dead or alive? In C. K. Georgiadis, P. Kefalas, and D. Stamatis, editors, *Local Proceedings of the Sixth Balkan Conference in Informatics, Thessaloniki, Greece, September 19-21, 2013*, volume 1036 of *CEUR Workshop Proceedings*, page 128. CEUR-WS.org, 2013.

[18] R. C. Martin. *Agile Software Development: Principles, Patterns, and Practices*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2003.

[19] J. McCarthy, M. I. of Technology. Computation Center, and M. I. of Technology. Research Laboratory of Electronics. *Lisp one five programmer's manual*. Massachusetts Institute of Technology, 1965.

[20] R. L. Nord, I. Ozkaya, and R. S. Sangwan. Making architecture visible to improve flow management in lean software development. *IEEE Software*, 29(5):33–39, 2012.

[21] L. J. Osterweil. Software processes are software too. In W. E. Riddle, R. M. Balzer, and K. Kishida, editors, *ICSE*, pages 2–13. ACM Press, 1987.

[22] S. P. Reiss. Dynamic detection and visualization of software phases. *ACM SIGSOFT Software Engineering Notes*, 30(4):1–6, 2005.

[23] W. Royce. Managing the development of large software systems: Concepts and techniques. In *WESCOM*, 1970.

[24] P. Spacek, C. Dony, and C. Tibermacine. A component-based meta-level architecture and prototypical implementation of a reflective component-based programming and modeling language. In *Proceedings of the 17th International ACM Sigsoft Symposium on Component-based Software Engineering*, CBSE '14, pages 13–22, New York, NY, USA, 2014. ACM.

[25] P. Spacek, C. Dony, C. Tibermacine, and L. Fabresse. Bridging the Gap between Component-based Design and Implementation with a Reflective Programming Language. Technical report, Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier - LIRMM , Unité de Recherche Informatique et Automatique - URIA, July 2013.

[26] P. Spacek, C. Dony, C. Tibermacine, and L. Fabresse. Wringing out objects for programming and modeling component-based systems. In *Proceedings of the Second International Workshop on Combined Object-Oriented Modelling and Programming Languages*, ECOOP'13, pages 2:1–2:6, New York, NY, USA, 2013. ACM.

[27] G. Spanoudakis and A. Zisman. Software traceability: a roadmap. *Handbook of Software Engineering and Knowledge Engineering*, 3:395–428, 2005.

[28] M. T. T. That, S. Sadou, and F. Oquendo. Using architectural patterns to define architectural decisions. In T. Männistö, A. M. Babar, C. E. Cuesta, and J. Savolainen, editors, *WICSA/ECSA*, pages 196–200. IEEE, 2012.

[29] C. Tibermacine, R. Fleurquin, and S. Sadou. A family of languages for architecture constraint specification. *Journal of Systems and Software*, 83(5):815–831, 2010.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 283–290
DOI: 10.15439/2016F485
ACSIS, Vol. 9. ISSN 2300-5963

From UML State Machines to code and back again!

Van Cam Pham, Ansgar Radermacher, Sébastien Gérard
CEA-List, Laboratory of Model-Driven Engineering for Embedded Systems (LISE)
Gif-sur-Yvette, France
Email: first-name.lastname@cea.fr

***Abstract*—UML state machines and their visual representations are much more suitable to describe logical behaviors of system entities than equivalent text based description such as IF-THEN-ELSE or SWITH-CASE constructions. Although many industrial tools and research prototypes can generate executable code from such a graphical language, generated code could be manually modified by programmers. After code modifications, round-trip engineering is needed to make the model and code consistent, which is a critical aspect to meet quality and performance constraints required for software systems. Unfortunately, current UML tools only support structural concepts for round-trip engineering such as those available from class diagrams. In this paper, we address the round-trip engineering of UML state-machine and its related generated code. We propose an approach consisting of a forward process which generates code by using transformation patterns, and a backward process which is based on code pattern detection to update the original state machine model from the modified code. We implemented a prototype and conducted several experiments on different aspects of the round-trip engineering to verify the proposed approach.**

I. Introduction

THE so-called Model-Driven Engineering (MDE) approach relies on two paradigms, abstraction and automation [1]. It is recognized as very efficient for dealing with complexity of today's systems. Abstraction provides simplified and focused views of a system and requires adequate graphical modeling languages such as Unified Modeling Language (UML). Even, if the latter is not the silver bullet for all software related concerns, it provides better support than text-based solutions for some concerns such as architecture and logical behavior of application development. UML state machines (USMs) and their visual representations are much more suitable to describe logical behaviors of system entities than any equivalent text based descriptions. The gap from USMs to system implementation is reduced by the ability of automatically generating code from USMs [2], [3], [4], [3].

Ideally, a full model-centric approach is preferred by MDE community due to its advantages [5]. However, in industrial practice, there is significant reticence [6] to adopt it. On one hand, programmers prefer to use the more familiar textual programming language. On the other hand, software architects, working at higher levels of abstraction, tend to favor the use of models, and therefore prefer graphical languages for describing the architecture of the system. The code modified by programmers and the model are then inconsistent. Round-trip engineering (RTE) [7] is proposed to synchronize different software artifacts, model and code in this case [8]. RTE enables actors (software architect and programmers) to freely move between different representations [8] and stay efficient with their favorite working environment.

Unfortunately, current industrial tools such as for instance Enterprise Architect [9] and IBM Rhapsody[10] only support structural concepts for RTE such as those available from class diagrams and code. Compared to RTE of class diagrams and code, RTE of USMs and code is non-trivial. It requires a semantical analysis of the source code, code pattern detection and mapping patterns into USM elements. This is a hard task, since mainstream programming languages such as C++ and JAVA do not have a trivial mapping between USM elements and source code statements.

For software development, one may wonder whether this RTE is doable. That is, why do the industrial tools not support the propagation of source code modifications back to original state machines? Several possible reasons to this lack are (1) the gap between USMs and code, (2) not every source code modification can be reverse engineered back to the original model, and (3) the penalty of using transformation patterns facilitating the reverse engineering that may not be the most efficient (e.g. a slightly larger memory overhead).

This paper addresses the RTE of USMs and object-oriented programming languages such as C++ and JAVA. The main idea is to utilize transformation patterns from USMs to source code that aggregates code segments associated with a USM element into source code methods/classes rather than scatters these segments in different places. Therefore, the reverse direction of the RTE can easily statically analyze the generated code by using code pattern detection and maps the code segments back to USM elements. Specifically, in the forward direction, we extend the double dispatch pattern presented in [11]. Traceability information is stored during the transformations. We implemented a prototype supporting RTE of state-machine and C++ code, and conducted several experiments on different aspects of the RTE to verify the proposed approach. To the best of our knowledge, our implementation is the first tool supporting RTE of SM and code.

To sum up, our contribution is as followings:

- An approach to round-tripping USMs and object-oriented code.
- A first tooling prototype supporting RTE of USMs and C++ code.
- An evaluation of the proposed approach including:
  - An automatic evaluation of the proposed RTE approach with the prototype.

– A lightweight evaluation of the semantic conformance of the runtime execution of generated code.

The remainder of this paper is organized as follows: Our proposed approach is detailed in Section II. The implementation of the prototype is described in Section III. Section IV reports our results of experimenting with the implementation and our approach. Section V shows related work. The conclusion and future work are presented in Section VI.

II. APPROACH

This section presents our RTE approach. At first, it sketches USM concepts supported by this study. The outline and the detail of the approach are presented afterward.

A. Scope

A USM describes the behavior of an active UML class which is called context class. A USM has a number of states and well-defined conditional transitions. A state is either an atomic state or a composite state that is composed of sub-states and has at most one active sub-state at a certain time. Transitions between states can be triggered by external or internal events. An action (effect) can also be activated by the trigger while transitioning from one state to another state. A state can have associated actions such as *entry/exit/doActivity* executed when the state is entered/exited or while it is active, respectively.

B. Approach outline

Our RTE approach is based on the double-dispatch pattern presented in [11] for mapping from USM to a set of classes with embedded code fragments. Fig. 1 shows the outline of our RTE approach consisting of a forward and a backward/reverse (engineering) process. In the forward process, a USM is transformed into UML classes in an intermediate model. The use of the intermediate model facilitates the transformation from the USM to code. Each class of the intermediate model contains attributes, operations and method bodies (block of text) associated with each implemented operation. The transformation utilizes several patterns which will be presented later.

When the source code is modified, a syntactic analysis process belonging to the backward transformation checks whether the modified code conforms to the USM semantics (see Subsection II-D3 for the detail of the analysis). This transformation takes as input the created intermediate model and the USM to update these models sequentially. While the forward process can generate code from hierarchical and concurrent USMs, the backward one only works for hierarchical machines excluding pseudo-states which are *history*, *join*, *fork*, *choice* and *junction*. These features are in future work.

C. From UML state machine to UML classes

This section describes the forward process which utilizes transformation patterns for states, transitions, and events to an intermediate model and code. We start by a simple USM

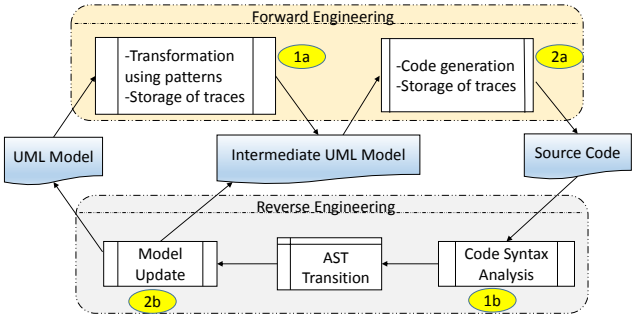


Fig. 1. Outline of state machine and code RTE

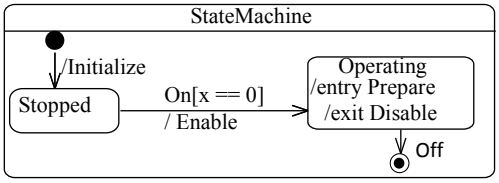


Fig. 2. An example of USM for tracing table

example in Fig. 2. It consists of two states (*Stopped* and *Operating*), two external events (*On* and *Off*), transitions, and an initial and a final pseudo state. Listing 1 shows a code portion generated from the USM following our approach.

Listing 1. A segment of C++ generated code

```
class CompositeState: public State {
protected:
  State* activeSubState;
public:
  bool dispatchEvent(Event* event) {
    bool ret = false;
    if (activeSubState != NULL) {
      ret = activeSubState->dispatchEvent(event);}
    return ret || event->processFrom(this);}}
StateMachine::StateMachine(Client* ctx){
  this->context = ctx;
  stopped = new Stopped(this, ctx);
  operating = new Operating(this, ctx);
  this->setIniDefaultState();
  this->activeSubState->entry();}
void StateMachine::setIniDefaultState(){
  this->context->Initialize();
  this->activeSubState = stopped;}
bool StateMachine::transition(Stopped* state,
      On* event) {
  if(this->context->guard(event)){
    this->activeSubState->exit();
    this->context->Enable(event);
    this->activeSubState = this->operating;
    this->activeSubState->entry();
    return true;}
  return false;}
bool StateMachine::transition(
    Operating* state, Off* event) {
  this->activeSubState->exit();
  //no action defined
  this->activeSubState = NULL;
  return true;}
class Stopped: public State {
private:
  StateMachine* ancestor;
public:
  virtual bool processEvent(On* event) {
    return ancestor->transition(this,event);}
}
class On: class Event {
public:
  processFrom(State* state) {
    state->processEvent(this);}
}
```

```
class Operating: public State {
47 private:
   StateMachine* ancestor;
49 public:
   void onEntryAction() {
51   context->Prepare();}
   void onExitAction() {
53   context->Disable();}
}
```

1) Transformation of states: Each state of the USM is transformed into a UML class. A state class inherits from a base class, namely, *State* (the detail of this class is not shown due to space limitation). *State* defines a reference to the context class, a process event operation for each event, state actions (*entry/exit/doActivity*). A bidirectional relationship is established between a state class and the state class associated with the containing state. For example, the USM example, considered as a composite state, has attributes typed by classes associated with its contained states, *Stopped* and *Operating* in Listing 1, lines 12-13. Inversely, attributes named *ancestor* (line 36) and typed by *StateMachine* in the classes *Stopped* and *Operating* are used to associate with the parent state.

A composite state class, which inherits from a base composite class (line 1), has an attribute *activeSubState* (line 3) indicating the active sub-state of the composite state and a *dispatchEvent* operation (line 5), which dispatches incoming events to the appropriate active state.

The *dispatchEvent* method implemented in the base composite state class delegates the incoming event processing to its active sub-state (line 8). If the event is not accepted by the active sub-state, the composite state processes it (line 9).

2) Transformation of events: Each event is transformed into a UML class (see lines 41-45 in Listing 1). An event can be either a *CallEvent*, *SignalEvent* or *TimeEvent* (see the UML specification for definitions of these events). An event class associated with a *CallEvent* inherits from a base event class and contains the parameters in form of attributes typed by the same types as those of the associated operation. The operation must be a member of the context class (a component as described above). For example, a call event *CallEventSend* associated with an operation named *Send*, which has two input parameters typed by *Integer*, is transformed into a class *CallEventSend* having two attributes typed by *Integer*. When a component receives an event, the event object is stored in an event queue.

A signal event enters the component through a port typed by the signal. The implementation view of this scenario depends on the mapping of component-based to object-oriented concepts. In the following, we choose the mapping done in Papyrus Designer [12]. In this mapping, the signal is transferred to the context class by an operation provided by the class at the associated port. Therefore, the transfer of a signal event becomes similar to that of *CallEvent*. For example, a signal event containing a data *SignalData* arrives at a port *p* of a component *C*. The transformation derives an interface *SignalDataInterface* existing as the provided interface of *p*. *SignalDataInterface* has only one operation *pushSignalData* whose body will be generated to push the event to the event queue of the component. Therefore, the processing of a *SignalEvent* is the same as that of a *CallEvent*. In the following sections, the paper only considers *CallEvent* and *TimeEvent*.

A *TimeEvent* is considered as an internal event. The source state class of a transition triggered by a *TimeEvent* executes a thread to check the expiration of the event duration as in [13] and puts the time event in the event queue of the context class.

3) Transformation of transitions and actions: Each action is transformed into an operation in the transformed context class. *Entry/Exit/doActivity* actions have no parameters while transition actions and guards accept the triggering event object. *doActivity* is implicitly called in the *State* class and executed in a thread which is interrupted when the state changes.

OnEntryAction and *OnExitAction* - abstract methods in the base state class *State* - are called by the entry and exit methods, respectively. Lines 50-53 in Listing 1 show how these methods are overwritten by the *Operating* class. *Prepare* and *Disable*, implemented in the context class, are called in these methods, respectively.

A transition is transformed into an operation taking as input the source state object and the event object similarly to DD. Transitions transformed from triggerless transition which has no triggering events accept only the source state object as a parameter. For example, the *Enable* action in the example is created in the context class and called by the transition method in lines 19-26. The guard *guard* is implemented as a method in the context class and called in line 21.

D. Reverse engineering from code to USM

This section describes the backward process.

1) Method Overall: The overall method for backward transformation is shown in Fig. 3. The modified code is first analyzed by partly inspecting the code syntax and semantics to guarantee that it is reversible. There are cases in which not all code modifications can be reversed back to the USM. The analysis also produces an output (*output2*) whose format is described later. If the intermediate model or the original USM is absent (the lower part of Fig. 3), a new intermediate model and a new USM are created from the UML model. In the contrary, the previous code taken, for instance, from control versioning systems is also semantically analyzed to have its output (*output1*) (the upper part of Fig. 3). *Output1* and *Output2* are then compared with each other to detect actual semantic changes which are about to be propagated to the original model.

Due to space limitation, we only show how to reconstruct (create) a new USM from the modified code.

2) Illustration example: To give an overview how the backward works, Fig. 4 presents a partition for mapping from the code segments generated from the example in Fig. 2 to actual USM concepts. Each partition consists of a code segment and the corresponding model element which are mapped in the backward direction. For example, the *Stopped* class in

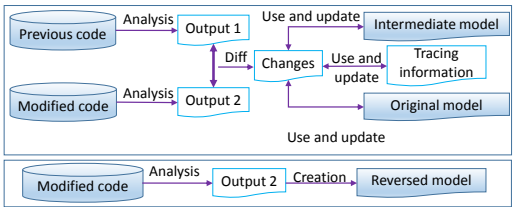


Fig. 3. Overall method for reversing code to state machine

code is mapped to the state *Stopped* of the state machine. The method *transition* is mapped back to the corresponding transition whose source state and triggering event are the input parameters' types of the method.

3) Semantic Analysis: The output of the semantic analysis contains a list of event names, a list of state names, a list of transitions in which each has a source state, a target state, a guard function, an action function and an event represented in so called abstract syntax tree (AST) transition [15]. For example, Fig. 5 presents the EMF [14] representation of transitions in a **C++** AST in which *IStructure* and *IFunctionDeclaration* represent a structure and a function in **C++**, respectively. Each state name is also associated with an ancestor state, an entry action, an exit action, a default sub-state and a final state. The output is taken by analyzing the AST. The analysis process consists of recognizing different patterns. Table I shows the main patterns including state, transition and event.

Algorithm 1 Semantic Analysis

Input: AST of code and a list of state classes stateList
Output: Output of semantic analysis

```
1: for s in stateList do
2:     for a in attribute list of s do
3:         if a and s match child parent pattern then
4:             put a and s into a state-to-ancestor map;
5:         end if
6:     end for
7:     for o in method list of s do
8:         if o is onEntryAction || o is onExitAction then
9:             analyzeEntryExit(o);
10:        else if o is processEvent then
11:            analyzeProcessEvent(o);
12:        else if o is setInitDefaultState & s is composite then
13:            analyzeInitDefaultState(s);
14:        else if o is timeout & s is a timedstate then
15:            analyzeTimeoutMethod(o);
16:            analyzeProcessEvent(s, o);
17:        end if
18:    end for
19: end for
```

Algorithm 1 shows the algorithm used for analyzing code semantics. Due to space limitation, *analyzeEntryExit*, *analyzeProcessEvent*, *analyzeInitDefaultState*, *analyzeTimeoutMethod* and *analyzeProcessEvent* are not presented but they basically follow the pattern description as above. In the first step of the analysis process, for each state class, it looks for an attribute typed by the state class, the class containing the attribute then becomes the ancestor class of the state class. The third steps checks whether the state class has an entry or exit action by looking for the implementation of the *onEntryAction* or *onExitAction*, respectively, in the state class to recognize the *Entry/Exit* action pattern. Consequently, event processing, initial default state of composite state and time event patterns are detected following the description as above.

4) Construction of USM from analysis output: If an intermediate model is not present, a new intermediate model and a new USM are created by a reverse engineering and transformation from the output of the analysis process. The construction is straightforward. At first, states are created. Secondly, UML transitions are built from the AST transition list. Lastly, action/guard/triggering event of a UML transition is created if the associated AST transition has these.

For example, assuming that we need to adjust the USM example shown in Fig. 2 by adding a guard to the transition from *Operating* to the final state. The adjustment can be done by either modifying the USM model or the generated code. In case of modifying code, the associated transition function in Listing 1 is edited by inserting an *if* statement which calls the guard method implemented in the context class. The change detection algorithm adds the transition function into the updated list since it finds that the source state, the target state and the event name of the transition is not changed. By using mapping information in the mapping table, the original transition in the USM is retrieved. The guard of the original transition is also created.

III. Implementation

The proposed approach is implemented in a prototype existing as an extension of the Papyrus modeler [15]. Each USM is created by using the state machine diagram implemented by Papyrus to describe the behavior of a UML class. Low-level USM actions are directly embedded in form of a block of code written in specific programming languages such as **C++/JAVA** into the USM. **C++** code is generated by the prototype but other object-oriented languages can be easily generated. The code generation consists of transforming the USM original containing the state machine to UML classes and eventually to code by a code generator following the proposed approach. The code generator can generate code for hierarchical and concurrent USMs. In the reverse direction, code pattern detection is implemented as described in the previous section to analyze USM semantics. If the generated code is modified, two options are supported by the prototype to make the USM and code consistent again. One is to create a new model containing the USM from the modified code in the same Eclipse project and the other one is to update the original USM by providing as input the intermediate model and the original model. At the writing moment, the prototype does not support the reverse of concurrent USMs and pseudo states, which are *history*, *join*, *fork*,*choice*, and *junction*.

IV. Experiments

In order to evaluate the proposed approach, we answer three questions stated as followings.

RQ1: A state machine *sm* is used for generating code. The generated code is reversed by the backward transformation to produce another state machine *sm'*. Are *sm* and *sm'* identical?

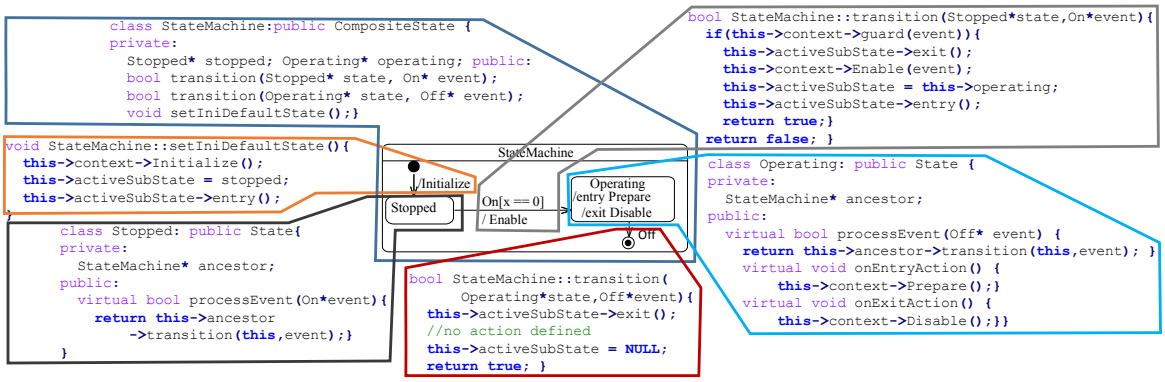


Fig. 4. USM element-code segment mapping partition

TABLE I
PATTERN RECOGNITION FOR REVERSE ENGINEERING GENERATED CODE

Pattern	Description
State	A state class inherits from the base state class or the composite base state class. For each state class, there must exist exactly one attribute typed by the state class inside another state class. The latter becomes the ancestor of the state class.
Composite state	A composite state class (CSC) inherits from the base composite state. For each sub-state the CSC has an attribute typed by the associated sub-state class. The CSC also implements a method named *setInitDefaultState* to set its default state. The CSC has a constructor is used for initializing all of its sub-state attributes at initializing time.
Entry action	If a state has an entry action, its associated state class implements *onEntryAction* that calls the corresponding action method implemented in the context class. Activity and exit patterns are recognized in the same way.
Event processing	If a state has an outgoing transition triggered by an event, the class associated with the state implements the *processEvent* method having only one parameter typed by the event class transformed from the event. The body calls the corresponding transition method of the ancestor class.
CallEvent	A call event class inherits from the base event class. The associated operation is found if the types of attributes of the event class match with the types of parameters of one method in the context class. A signal event is treated as a *CallEvent* as previously described.
TimeEvent	A transition is triggered by a *TimeEvent* if the state class associated with its source state implements the timed interface. The duration of the time event is detected in the transition method whose name is formulated as *"transition" + duration*.
Transition	Transition methods are implemented in the ancestor class, which is the class associated with the composite state owning the source state of the transition. The first parameter of the methods is the class representing the source state. The second parameter is the triggering event. Methods associated with *triggerless* transitions do not have a second parameter. The body of external and internal transition methods contains ordered statements including exiting the source state, executing transition action (effect), changing the active state to the target or null if the target is the final state, and entering the changed active state by calling entry. The body can have an if statement to check the guard of the transition.
Effect/ guard	Transition actions and guards are implemented in the context class.

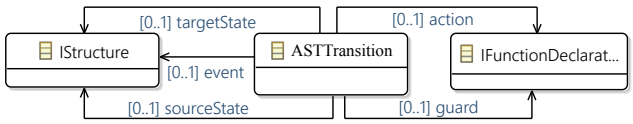


Fig. 5. Transitions output from the analysis

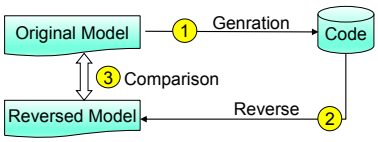


Fig. 6. Evaluation methodology to answer RQ1

In other words: whether the code generated from USMs model can be used for reconstructing the original model. This question is related to the *GETPUT* law defined in [16].

RQ2: **RQ1** is related to the static aspect of generated code. **RQ2** targets to the dynamic aspect. In other words, whether the runtime execution of code generated from USMs by the generator is semantic-conformant [17]?

This section reports our experiments targeting the three questions. Two types of experiments are conducted and presented in Subsections IV-A and IV-B, respectively.

A. Reversing generated code

This experiment is targeting **RQ1**. Fig. 6 shows the experimental methodology to answer **RQ1**. The procedure for this experiment, for each original UML model containing a state machine, consists of 3 steps: (1) code is generated from an **original model**, (2) the generated code is reversed to a **reversed model**, and (3) the latter is then compared to the **original state machine**.

Random models containing hierarchical state machines are automatically generated by a configurable model generator. The generator can be configured to generate a desired average number of states and transitions. For each model, a context

TABLE II
THREE OF MODEL RESULTS OF GENERATION AND REVERSE: ABBREVIATIONS ARE ATOMIC STATES (AS), COMPOSITE STATES (CS), TRANSITIONS (T), CALL EVENTS (CE), TIME EVENTS (TE)

Test ID	AS	CS	T	CE	TE	Is reverse correct?
1	47	33	234	145	40	Yes
2	42	38	239	145	36	Yes
..	..	..	..	..	..	Yes
300	41	39	240	142	37	Yes

class and its behavior described by a USM are generated. Each USM contains 80 states including atomic and composite states, more than 234 transitions. The number of lines of generated C++ code for each machine is around 13500. Names of the generated states are different. An initial pseudo state and a final state are generated for each composite state and containing state machine. Other elements such as call events, time events, transition/entry/exit actions and guards are generated with a desired configuration. For each generated call event, an operation is generated in the context class which is also generated. The duration is generated for each time event.

Table II shows the number of several types of elements in the generated models, including the comparison results, for 3 of the 300 models created by the generator. We limited ourselves to 300 models for practical reasons. No differences were found during model comparison. The results of this experiment show that the proposed approach and the implementation can successfully do code generation from state machines and reverse.

B. Semantic conformance of runtime execution

a) Bisimulation for semantic-conformance: To evaluate the semantic conformance of runtime execution of generated code, we use a set of examples provided by Moka [18]. Moka is a model execution engine offering Precise Semantics of UML Composite Structures [17]. Fig. 7 shows our method. We first use our code generator to generate code (Step (1)) from the Moka example set. Step (2) simulates the examples by using Moka to extract the sequence (*SimTraces*) of observed traces including executed actions. The sequence (*RTTraces*) of traces is also obtained by the runtime execution of the code generated from the same state machine in a Step (3). The generated code is semantic-conformant if the sequences of traces are the same for both of the state machine and generated code [19]. The current version of Moka does not support simulation for *TimeEvent* and history pseudo states, we therefore leave experiments for *TimeEvent* as future work.

For example, Fig. 8 (a) shows a USM example with triggerless transitions (*autotransitions*) *T3*. The USM contains two states, *Waiting*, which is the initial state, and *Incrementing*, which increases an integer number from 0 to 5 by using the effect of *T3*. The latter also has a guard checking whether the number is less than 5. The increase is executed after the USM receives an event named *start* to transition the initial state *Waiting* to *Incrementing*. Suppose that executions of the effects of *T3* and *T4* produce traces <*T3*> and <*T4*> (by using

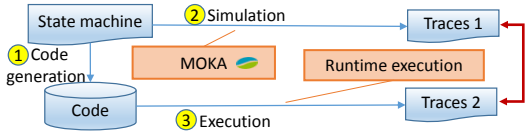


Fig. 7. Semantic conformance evaluation methodology

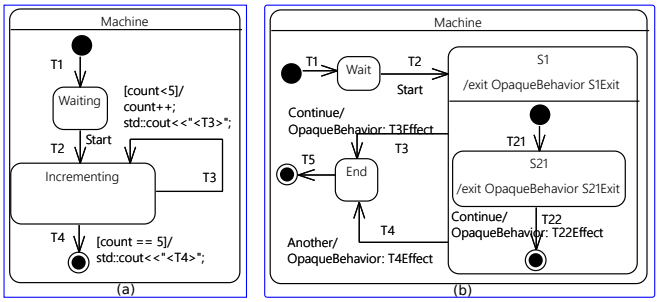


Fig. 8. Self-triggerless transition and event processing example

MOKA, e.g.), respectively. Due to the guard of *T3*, the effect of *T3* is executed five times followed by an execution of the effect of *T4*. After the completion of the USM, the obtained sequence of traces is <*T3*><*T3*><*T3*><*T3*><*T3*><*T4*> (since the *Incrementing* state does not have an *entry*, *exit*, or a *doActivity*, only the transition effect *T3* produces traces). The sequence *RTTraces* obtained by the runtime execution must be equivalent. *RTTraces* is obtained by simply printing logging information for each action (effect).

Within our scope as previously defined 30 examples of the Moka example set are tested. *SimTraces* and *RTTraces* for each case are the same. This indicates that, within our study scope, the runtime execution of code generated by our generator can produce traces semantically equivalent to those obtained via simulation.

After experimenting with our code generator, we compare our results to the observed traces obtained by executing code generated Umple [20]. We find that the obtained traces in case of Umple are not UML-compliant in triggerless transitions and some cases of event processing. Specifically, for the example in Fig. 8 (a), code generated by Umple only produces <*T3*> as the trace sequence. Umple does not support events which are accepted by sub-states and the corresponding composite state as in Fig. 8 (b) in which both *S1* and *S21* accept the event *Continue*. As the processing event example in Fig. 8 (b), assuming that there is an event *Continue* incoming to the state machine which has a current configuration *(S1, S21)* as current active states. While, according to the UML specification, the incoming event should be processed by the inner states of the active composite/concurrent state if the inner states accept it, otherwise the parent state does. Therefore, the next configuration should be *(S1, final state)* and the *T22Effect* effect of the transition *T22* should be executed.

b) Finite state machine: In addition to the experiment using MOKA, we evaluate the semantic-conformance by using deterministic finite state machines (FSMs). The latter is a

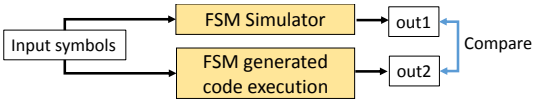


Fig. 9. FSM experiment method

mathematical model of computation and also a simplified version of UML state machine. In this experiment, we use FSMs for recognizing input symbols. Each FSM contains many atomic states. The active state of the FSM can be changed following the acceptance of an input symbol. Fig. 9 shows our method to experiment. For each FSM, we create an equivalent USM. Each input symbol of the FSM is considered as an event of the USM. We use the FSM simulator in [21] to generate and simulate FSMs. For each FSM, a list of observed states is recorded as output (*out1*) of the simulation for each symbol list. The latter is also the input of the generated code runtime execution of the equivalent USM which produces an output *out2*. We then compare *out1* and *out2*.

We limit the number of FSMs to 20 and the number of symbol list for each FSM to 30 for practical concerns. 600 sequences of states obtained by the simulation and a same number of sequences taken by the runtime execution are respectively compared and found being equal. This results that our code generation approach can produce semantic-conformant code in case of FSM.

V. Related Work

Our work is motivated by the desire to reduce the gap between and synchronize artifacts at different levels of abstraction, model and code in particular, in developing reactive systems. Specifically, the usage of USMs for describing the logical behavior of complex, reactive systems is indispensable. In the following sections, we compare our approach to related topics recorded in the literature.

A. Implementation and code generation techniques for USMs

Implementation and code generation techniques for USMs are closely related to the forward engineering of our RTE.

Switch/if is the most intuitive technique implementing a "flat" state machine. The latter can be implemented by either using a scalar variable [2] and a method for each event or using two variables as the current active state and the incoming event used as the discriminators of an outer switch statement to select between states and an inner one/if statement, respectively. The double dimensional state table approach [3] uses one dimension represents states and the other one all possible events. The behavior code of these techniques is put in one file or class. This practice makes code cumbersome, complex, difficult to read and less explicit when the number of states grows or the state machine is hierarchical. Furthermore, these approaches requires every transition must be triggered by at least an event. This is obviously only applied to a small sub-set of USMs.

State pattern [4], [3] is an object-oriented way to implement flat state machines. Each state is represented as a class and each event as a method. Separation of states in classes makes the code more readable and maintainable. This pattern is extended in [22] to support hierarchical-concurrent USMs. However, the maintenance of the code generated by this approach is not trivial since it requires many small changes in different places.

Many tools, such as [10], [9], apply these approaches to generate code from USMs. Readers of this paper are recommended referring to [23] for a systematic survey on different tools and approaches generating code from USMs.

Double-dispatch (DD) pattern in [11] in which represent states and events as classes. Our generation approach relies on and extends this approach. The latter profits the polymorphism of object-oriented languages. However, DD does not deal with triggerless transitions and different event types supported by UML such as *CallEvent*, *TimeEvent* and *SignalEvent*. Furthermore, DD is not a code generation approach but an approach to manually implementing state machines.

B. Round-trip engineering

Our RTE is related to synchronization of model-code and models themselves that a large number of approaches support. This paper only shows the most related approaches.

Model-code synchronization

Commercial and open-source tools such as [9], [10] only support RTE of architectural model elements and code. Systematic reviews of some of these tools are available in [24].

Some RTE techniques restrict the development artifact to avoid synchronization problems. Partial RTE and protected regions [25] aim to preserve code changes which cannot be propagated to models. This approach separates the code regions that are generated from models from regions which are allowed to be edited by developers. This form of RTE is unidirectional only and does not support iterative development [26] Our approach does not separate different regions but supports a semantic code analysis in the reverse direction.

Fujaba [27] offers an RTE environment. An interesting part of Fujaba is that it abstracts from Java source code to UML class diagrams and a so-called story-diagrams. Java code can also be generated from these diagrams. RTE of these diagrams and code works but limited to the naming conventions and implementation concepts of Fujaba which are not UML-compliant.

Model synchronization

RTE of models is tackled by many approaches categorized by its model transformation from total, injective, bi-directional to partial non-injective transformations [7]. Techniques and technologies, such as Triple Graph Grammar (TGG) [28] and QVT-Relation [29], allow synchronization between source and target elements that have non-injective mappings. These techniques require a mapping model to connect the source and target models which need to be persisted in a model store

[30]. Mappings between USMs and code in our approach are non-bijective. We only use simple tables for storing tracing information.

Differently from other approaches, ours is specific to RTE of USMs and code. The goal is to provide a full model-code synchronization supporting for rapidly, iteratively, and efficiently developing reactive systems.

VI. Conclusion

This paper presented a novel approach to RTE from USMs to code and back. The forward process of the approach is based on different patterns transforming USM elements into an intermediate model containing UML classes. Object-oriented code is then generated from the intermediate model by existing code generators. In the backward direction, code is analyzed and transformed into an intermediate whose format is close to the semantics of USMs. USMs are then reconstructed from the intermediate format.

The paper also showed the results of several experiments on different aspects of the proposed RTE with the tooling prototype. Specifically, the experiments on the correctness and semantic conformance of code using the proposed RTE are conducted. Although, the reverse direction only works if manual code is written following pre-defined patterns, the semantics of USMs is explicitly present in generated code and easy to follow/modify.

While the semantic conformance of code generated is critical, the paper only showed a lightweight experiment on this aspect. A systematic evaluation is therefore in future work. We will also compare our code generation approach with commercial tools such as Rhapsody and Enterprise Architect. Furthermore, as evaluated in [7], the approach inheriting from the double-dispatch trades a reversible mapping for a slightly larger overhead. For the moment, the approach does not support RTE of concurrent state machines and several pseudo-states. Hence, future work should resolve these issues.

Acknowledgment

The work presented in this paper is supported by the European project SafeAdapt, grant agreement No. 608945, see *http://www.SafeAdapt.eu*. The project deals with adaptive system with additional safety and real time constraints. The adaptation and safety aspects are stored in different artifacts in order to achieve a separation of concerns. These artifacts need to be synchronized.

References

[1] G. Mussbacher, D. Amyot, R. Breu, J.-m. Bruel, B. H. C. Cheng, P. Collet, B. Combemale, R. B. France, R. Heldal, J. Hill, J. Kienzle, and M. Schöttle, "The Relevance of Model-Driven Engineering Thirty Years from Now," *ACM/IEEE 17th International Conference on Model Driven Engineering Languages and Systems (MODELS)*, pp. 183–200, 2014.

[2] G. Booch, J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide*, 1998, vol. 3.

[3] B. P. Douglass, *Real-time UML : developing efficient objects for embedded systems*, 1999.

[4] A. Shalyto and N. Shamgunov, "State machine design pattern," *Proc. of the 4th International Conference on.NET Technologies*, 2006.

[5] B. Selic, "What will it take? a view on adoption of model-based methods in practice," *Software & Systems Modeling*, vol. 11, no. 4, pp. 513–526, 2012.

[6] J. Hutchinson, M. Rouncefield, and J. Whittle, "Model-driven engineering practices in industry," in *Proceedings of the 33rd International Conference on Software Engineering*, ser. ICSE '11. New York, NY, USA: ACM, 2011, pp. 633–642.

[7] T. Hettel, M. Lawley, and K. Raymond, "Model synchronisation: Definitions for round-trip engineering," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5063 LNCS, 2008, pp. 31–45.

[8] S. Sendall and J. Küster, "Taming model round-trip engineering," *Proceedings of Workshop Best Practices for Model-Driven Software Development*, p. 13, 2004.

[9] SparxSystems, "Enterprise Architect," Sep. 2016. [Online]. Available: http://www.sparxsystems.eu/start/home/

[10] IBM, "Ibm Rhapsody." [Online]. Available: http://www.ibm.com/developerworks/downloads/r/rhapsodydeveloper/

[11] V. Spinke, "An object-oriented implementation of concurrent and hierarchical state machines," *Information and Software Technology*, vol. 55, no. 10, pp. 1726–1740, Oct. 2013.

[12] "Papyrus Designer." [Online]. Available: https://wiki.eclipse.org/Papyrus_Designer

[13] I. Niaz and J. Tanaka, "Mapping UML statecharts to java code." *IASTED Conf. on Software Engineering*, pp. 111–116, 2004.

[14] R. Gronback, "Eclipse Modeling Project." [Online]. Available: http://www.eclipse.org/modeling/emf/

[15] CEA-List, "Papyrus Homepage Website," https://eclipse.org/papyrus/.

[16] J. N. Foster, M. B. Greenwald, J. T. Moore, B. C. Pierce, and A. Schmitt, "Combinators for Bidirectional Tree Transformations: A Linguistic Approach to the View-update Problem," *ACM Trans. Program. Lang. Syst.*, vol. 29, no. 3, May 2007.

[17] OMG, "Precise Semantics Of UML Composite Structures," no. October, 2015.

[18] "Moka Model Execution." [Online]. Available: https://wiki.eclipse.org/Papyrus/UserGuide/ModelExecution

[19] J. O. Blech and S. Glesner, "Formal verification of java code generation from uml models," in *... of the 3rd International Fujaba Days*, 2005, pp. 49–56.

[20] O. Badreddin, T. C. Lethbridge, A. Forward, M. Elasaar, and H. Aljamaan, "Enhanced Code Generation from UML Composite State Machines," *Modelsward 2014*, pp. 1–11, 2014.

[21] F. Simulator, "FSM Simulator," http://ivanzuzak.info/noam/webapps/fsm_simulator/.

[22] I. A. Niaz, J. Tanaka, and others, "Mapping UML statecharts to java code." in *IASTED Conf. on Software Engineering*, 2004, pp. 111–116.

[23] E. Domínguez, B. Pérez, A. L. Rubio, and M. A. Zapata, "A systematic review of code generation proposals from state machine specifications," pp. 1045–1066, 2012.

[24] D. Cutting and J. Noppen, "An Extensible Benchmark and Tooling for Comparing Reverse Engineering Approaches," *International Journal on Advances in Software*, vol. 8, no. 1, pp. 115–124, 2015. [Online]. Available: https://ueaeprints.uea.ac.uk/53612/

[25] D. Frankel, *Model Driven Architecture: Applying MDA to Enterprise Computing*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[26] S. Jörges, "Construction and evolution of code generators: A model-driven and service-oriented approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 1–265, 2013.

[27] T. Kleín, U. A. Nickel, J. Niere, and A. Zündorf, "From uml to java and back again," University of Paderborn, Paderborn, Germany, Tech. Rep. tr-ri-00-216, September 1999.

[28] H. Giese and R. Wagner, "Incremental model synchronization with triple graph grammars," in *Model Driven Engineering Languages and Systems*. Springer, 2006, pp. 543–557.

[29] Q. Omg, "Meta Object Facility (MOF) 2 . 0 Query / View / Transformation Specification," *Transformation*, no. January, pp. 1–230, 2008.

[30] G. Bergmann, I. Ráth, G. Varró, and D. Varró, "Change-driven model transformations: Change (in) the rule to rule the change," *Software and Systems Modeling*, vol. 11, no. 3, pp. 431–461, 2012.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 291–298
DOI: 10.15439/2016F362
ACSIS, Vol. 9. ISSN 2300-5963

Competencies outside Agile Teams' Borders: The Extended Scrum Team

Gerard Wagenaar
Avans University of Applied Science, Academy for Engineering & ICT, Lovensdijkstraat 61, 4818 AJ Breda, the Netherlands
Email: g.wagenaar@avans.nl

Sietse Overbeek
Utrecht University, Faculty of Science, Department of Information and Computing Sciences, Princetonplein 5, 3584 CC Utrecht, the Netherlands
Email: s.j.overbeek@uu.nl

Remko Helms
Open University, Faculty of Management, Science and Technology, Valkenburgerweg 177, 6419 AT Heerlen, the Netherlands
Email: remko.helms@ou.nl

***Abstract*—According to the Scrum process framework a Scrum team should have all necessary competencies to accomplish its work. Fragmented and anecdotal evidence hints at Scrum teams still needing additional, external competencies. To contribute to theories on Scrum team composition and practitioner's concerns in staffing a Scrum team we investigated Scrum teams' cross-functionality: To whom do Scrum teams turn for additional competencies, which competencies are involved and how are Scrum teams aware of additional competencies they need? To this extent we analysed the communication in three Scrum teams during one of their Sprints. Our results show that additional competencies are called for, not only on an ad hoc basis, but also on a structural basis. To include those structural competencies the notion of an extended Scrum team is introduced.**

I. Introduction

THE application of an agile software development method, for instance XP or Scrum, is nowadays common in delivering state-of-the-art software [1-2]. Team members with high competence and expertise are a critical success factor in agile software development, especially with regard to timeliness and cost [3].

One of the guidelines accompanying Scrum as a framework for developing and sustaining complex (software) products states on team composition: "*Cross-functional teams have all competencies needed to accomplish the work without depending on others not part of the team*" [4, *p.4*]. Such a cross-functional team should, among others, have skills with regard to software analysis, design and coding [5].

In general, although often implicit, membership of a team is considered to be full time during a Scrum project, certainly for members of the Development Team [6-7]. Emergent team members may support a team during software development [8] or even stronger it is often not clear which members belong to a team and which do not [9] or: "*team boundaries are often permeable*" [10, *p.157*]. Part time membership of a Scrum team could therefore be an option to consider [5], [11].

Fragmented and anecdotal evidence already hints at Scrum teams still needing additional, external competencies [12-14], but this evidence was gathered as a by-product of more general research on communication in agile projects. Unlike this research we only focus on composition of Scrum teams, especially their cross-functionality. We determine its boundaries and look for additional competencies not included in the team. To this extent only we analyse the communication within and outside team boundaries of Scrum teams' members to identify which additional competencies they require and on which basis. In this way our research on the one hand contributes to a better understanding of team composition in Scrum software development, and especially in the competencies included in the team, and those outside its boundaries, and on the other hand allows practitioners to mirror their way of working, and support them in the formation of a Scrum team.

The remainder of this paper is organised as follows. In section 2 we outline the theoretical background relating to our work, based on a literature review. In section 3 we present our research method. The results for three case studies are presented in section 4, followed by a discussion in section 5. Section 6 is the final section in which our conclusions are presented, with their limitations and future work.

II. Theoretical background

In terms of socio-technical congruence [15] the Scrum process framework [4] describes the coordination requirements established by the dependencies among tasks. Upon inspection of the actual coordination activities a match between coordination requirements and activities may be established; such a match has proven to be beneficial to several aspects of software development, for instance reducing the resolution time of modification requests [15] or software build success [16].

We apply socio-technical congruence specifically to the use of competencies in a Scrum team. The coordination requirements defined by the Scrum process framework state cross-functionality. Should a perfect match with the actual activities exist, i.e. they can be carried out without external competencies, then a Scrum team is indeed cross-functional.

Competencies for agile software development have been divided into three major categories, forming a pyramid of agile competencies with engineering practices at the bottom via management practices to agile values at the top [17] or, similarly, technical skills, people or soft skills, and attitudes [18]. Support for the two latter categories is a responsibility

of a Scrum master, although an agile coach, operating outside a team's boundaries, is also often involved [19], [20]. Since these are individual rather than team competencies, we use socio-technical congruence to focus on the former category: Engineering skills.

There are already indications that Scrum teams are not entirely cross-functional in this respect. No matter how tightly-knit agile teams are, they need to interact with roles outside the team (e.g. user experience designers, database administrators, system testers), because in practice it is infeasible for all individuals with relevant expertise to be part of the team [14]. This is, for instance, confirmed in the communication between the disciplines of agile development and user experience design [21].

Although internally distributing expertise in agile teams ultimately leads to successful cross-functional teams [22], this appears to be an ideal situation. At least until then, but perhaps even on a more continuous basis, it is an unrealistic goal to aspire to include all competencies in a Scrum development team itself, with its 3 to 9 members. Additional roles, contributing competencies to the Scrum team, have already been hinted at; they could be user experience designers, database administrators, system testers [14], acceptance testers, user interaction designer, or technical writers [13]. Part time members could also be system or database administrators [5].

Four ways to locate expertise in a Scrum team itself have been revealed:

1. Communicating frequently,
2. Working closely together,
3. Declaring self-identified expertise in order to let others know what a member can contribute to the team,
4. Using an expertise directory [23].

It could be the case that these also apply to locating expertise outside a team's border, but this has not been established yet.

Coordinating expertise outside agile teams benefits from five factors:

1. Availability refers to the ability of external specialists to be present in agile teams when their expertise is needed,
2. Agile mind-set concerns dealing with external specialists who might be unfamiliar with agile methods and thus introducing problems for the team, for instance external specialists not being able to align work with the sprints,
3. Stability refers to keeping agile teams stable with a low rate of team members and external specialists turnover,
4. Knowledge retention involves capturing external specialists' knowledge and preserving the knowledge in agile teams,
5. Effective communication is defined as the activity of conveying sufficient information between agile teams and external specialists [24].

These factors were established from responses from individuals, which were not necessarily joined in teams. Although important factors in coordinating expertise outside an agile team were identified, it does not answer the question how this expertise is identified in the first place or what expertise is involved.

Software development in general needs expertise finding to facilitate unplanned collaborative work among software developers, but: "*Who are these additional people, and why are they contributing, or why have they been contacted …?*" [8, *p. 89*]. And these questions are equally applicable to the use of Scrum in software development.

To identify partners outside Scrum team boundaries, it is important to realize that both Product Owner and Scrum Master act as linking pins with the team's environment; the Product Owner in managing the Product Backlog in cooperation with stakeholders, the Scrum Master in serving the (outside) organization, for instance in helping employees and stakeholders understand and enact Scrum and empirical product development. Communication between the Product Owner and the Scrum Master on the one hand and partners like management and stakeholders on the other hand are thus already included in the Scrum process framework. Taking this into account and under the assumption that the Scrum team members attend all regular Scrum events we present an initial sketch of a Scrum team and its external partners (Fig. 1); this sketch can thus be considered as the common composition of a Scrum team (and its partners).

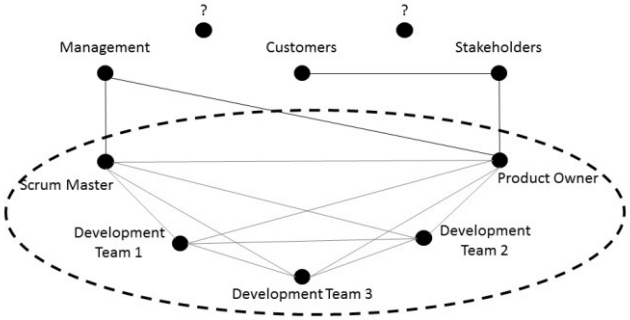


Fig. 1. Communication inside and outside Scrum teams

Question marks represent yet unidentified partners with additional competencies. The members Development Team 1 – 3 represent team members with competencies with regard to software analysis, design and coding [5], where each node represents several team members, depending on the team size. We will use this model both to guide our data analysis as well as to compare its contents with our results.

III. Case Study Design

To investigate the cross-functionality of Scrum teams our research analyses the communication of Scrum teams outside their team boundaries to identify which additional competencies they require and on which basis. To allow for rich evidence we selected an exploratory comparative case study approach as our research method with as unit of analysis one sprint of Scrum software development. This approach is an accustomed way to investigate phenomena in a context where events cannot be controlled and where the focus is on contemporary events [25]. For this case study we drew up the following research questions:

RQ1 To whom do Scrum teams turn to for additional competencies and which competencies are involved?
RQ2 How are Scrum teams aware of what additional competencies are needed?

We used a protocol to guide the case study [26]. This protocol contained:

- Its purpose, guidelines for data and document storage, and publication.
- A brief overview of the case research method.
- Detailed procedures for conducting each case, to ensure uniformity in the data collection process and consequently facilitate both within and cross case analyses.
- Research instruments.
- Guidelines for data analysis.

Given our goal and research questions, organizations were required to use Scrum as software development method with a team of at least 5 members; we chose this lower limit to allow us to find indeed cross-functional teams. We approached three organizations to participate in our research and all three were willing to do so; they all had a team size between 5 and 10 members. We name them Controller, Sunflower and Local for reasons of confidentiality.

A. Data collection

The primary data collection method was semi-structured interviewing of team members. We also inspected available documents and/or the contents of information systems which, in some cases, were used to support the Scrum process. Our questionnaire addressed communication both between team members, exemplified in the use of Scrum practices, as well as communication between team members and non-team members. Some examples of questions of the latter category (from the protocol) are:

- With whom did you communicate (mainly) during the sprint? Please mention all people and include people who might have been outside the scope of the sprint (in first instance).
- What was the communication about?

Interviews lasted, on average, 60 minutes. In total approximately 12 hours with 13 interviewees were available. To achieve a representative sample a team's Scrum Master and Product Owner were always included and, depending on the size of the team, 2 to 4 developers, including designers and testers when these roles were explicitly assigned in a team. Six more additional interviews were held either before the other interviews started to provide general context, or afterwards, to clarify remaining issues.

Thirteen interviews were transcribed and coded, with a combination of open and axial coding [27]. We also used our preliminary sketch (Fig. 1) to guide the coding; it guided the coding in the sense that external partners were either classified as manager, customer or stakeholder or an additional partner not yet included (previously a question mark). Summaries of interviews were consulted with the interviewees. For each organization the results of the interviews and additional material were bundled in a case study report.

B. Validity

Validity of our research method depends on four widely used criteria: construct validity, internal and external validity, and reliability [25].

Construct validity identifies operational measures for the concepts under study. To enhance construct validity (1) key informants should review draft case study reports, (2) multiple sources of evidence should be used, and (3) a chain of evidence should be established [25]. We applied all three: (1) Each interviewee was provided with a summary report of the interview and key interviewees commented on a draft case study report, (2) various team members were involved to complement viewpoints, and (3) interviews (and other materials) were linked to conclusions by using the tool NVivo; NVivo is a software package to aid qualitative data analysis (www.qsrinternational.com).

Internal validity is mainly a concern for explanatory case studies [25], [28]. Our case study is exploratory, but we did apply pattern matching, one of the analytical techniques recommended to enhance internal validity, by consistently coding our base material.

External validity defines the domain to which a case study's findings can be generalized. The use of replication logic is listed as the main guarantee for this [25]. Using a multiple-case study on the basis of a common questionnaire contributes to external validity, and thus to generalizability of results.

Reliability should demonstrate that the study can be repeated. The use of a case study protocol and the development of a case study database [25] were both applied in our study to increase reliability.

IV. Results

In this section we describe the results from our case studies. We first give an impression of the organizations and the composition of their Scrum teams (Table 1). For Controller the role of Scrum Master coincides with one of the developers/testers; for Sunflower the role of Scrum Master coincides with one of the senior developers.

TABLE I. DESCRIPTION OF ORGANIZATIONS

Organization	Domain	Team composition			Scrum experience
		Product Owner	Development Team	Scrum Master	
Controller	Object management	1	2 developers 2 developers/testers	1	2 years
Sunflower	Floral industry	3	3 senior developers 1 junior developer	1	2½ years
Local	Government taxing	1	2 designers 4 developers 2 testers	1	1½ years

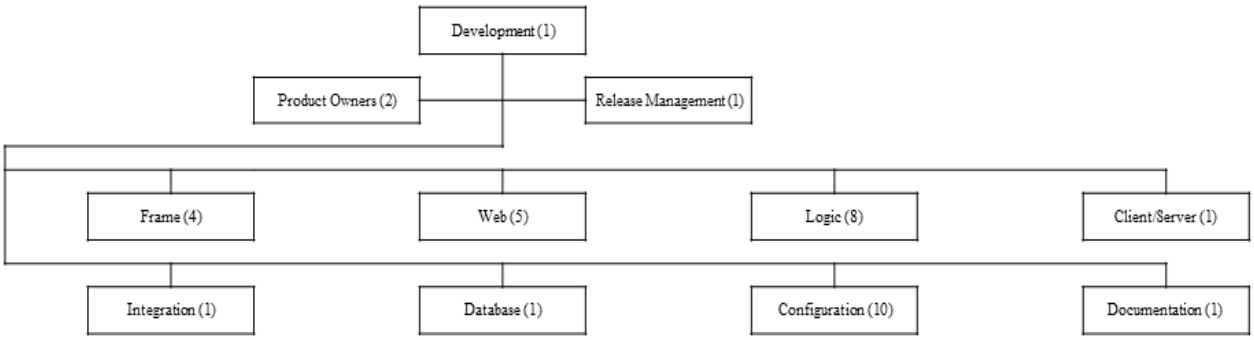


Fig 2 Controller organization chart for Development department

In the next three paragraphs we describe, for each organization in turn, the communication within, but mainly outside their Scrum teams' borders.

A. Controller

The Scrum team of Controller operates in a larger organizational context: the Development department with 35 employees (Fig. 2).The numbers in Fig. 2 refer to the number of employees in a group. The Frame group is responsible for the building blocks (basic components) of the software; Web adds a graphical skin. Together these groups produce a kind of half-fabricate. The groups Logic and Configuration build end products on the basis of these (supporting) components; the Scrum team is staffed from these two groups and complemented with one of the two Product Owners. In this Logic takes care of the process flow in the software, whereas Configuration is concerned with visual aspects, screens, buttons, reports, et cetera. Configuration is also responsible for testing the software. All other groups are small; they support the four groups mentioned before.

The Scrum team communicates in the framework of Scrum events. Sprint Planning Meeting, Daily Scrum, Sprint Review Meeting and Sprint Retrospective are all regularly scheduled 'meetings'. But team members also communicate with non-team members. The figure below (Fig. 3) demonstrates communication within and beyond the team's borders, where the customer is the only partner outside the organization Controller directly communicating with a team member.

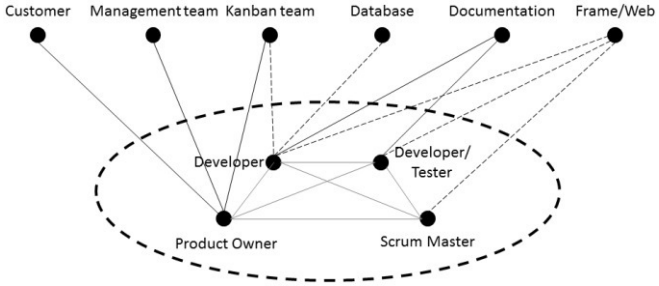


Fig. 3 Internal & external communication for Controller's Scrum team

The members of the Scrum team are shown in the lower half of Fig. 3; there is agreement on the team composition. In Fig. 3 only roles of team members are shown; the team in fact has 2 developers and 2 developers/testers (refer to Table 1). Apart from the Product Owner, all members are full time members.

External partners are shown in the upper half of Fig. 3. Partners are only involved in the team's activities part time, where in vast majority part time means less than a small number of hours per Sprint. Solid lines represent structural communication between the team members, where we defined structural as communication not only in the Sprint under consideration, but also in a majority of Sprints; dotted lines indicate ad hoc communication taking place in this Sprint only, but not necessarily in other Sprints. Furthermore we include only partners who are directly communicating with at least one team member as the team is the focus of our research.

Considering the external partners we focused on the unidentified partners (Fig. 1). Already identified partners for Controller then are Customer and Management team.

The Scrum team is only involved in the development of new parts of the software; maintenance is done by another team with Kanban. The Product Owner is responsible for coordination of the Kanban team; he - every employee is indicated as 'he', whether male or female - coordinates (the prioritization of) maintenance. It is his task to recognize overlapping activities of the Scrum and the Kanban team, with regard to components of the software, and to bring members of the two teams together whenever necessary. Although overlap does not occur every Sprint, communication is frequent whenever the two teams do work on the same piece of code to coordinate activities with regard to new code or adaptations to existing code.

The same applies for communication with the database specialist. He is consulted whenever sprint backlog items have impact on the database structure.

Documentation consists of a technical writer. He is responsible for the construction of a user guide and/or release notes in every Sprint and in cooperation with Development Team members. Communication is mainly on his initiative, where he has basic information available through a registration system.

Frame/Web, as producers of half-fabricates, are often contacted on details of the code they manufactured, although not in every Sprint.

The remaining groups in the department, Client/Server and Integration, were not mentioned to participate in the team's communication.

B. Sunflower

Sunflower belongs to the Small & Medium Enterprises (SME), more specifically a small company with a number of employees around 15; its organization chart is shown in Fig. 4, the numbers referring to the number of employees.

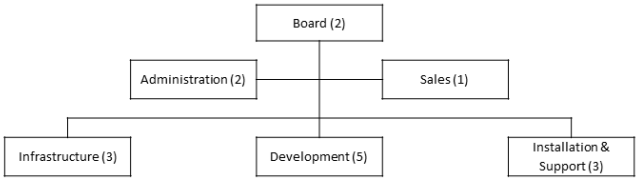


Fig. 4 Sunflower organization chart

Consultants, from Installation and Support, collectively function as Product Owner. The Development team consists of members of the group Development; the role of Scrum Master rests with a senior developer.

Members of the Scrum team communicate in the framework of Scrum. Sprint Planning Meeting and Daily Scrum are both regularly scheduled 'meetings', but they are skipped occasionally. Neither a Sprint Review Meeting nor a Sprint Retrospective is used by the team. Communication within and beyond the team's borders involves a restricted number of partners (Fig. 5), where customer is the only partner outside Sunflower.

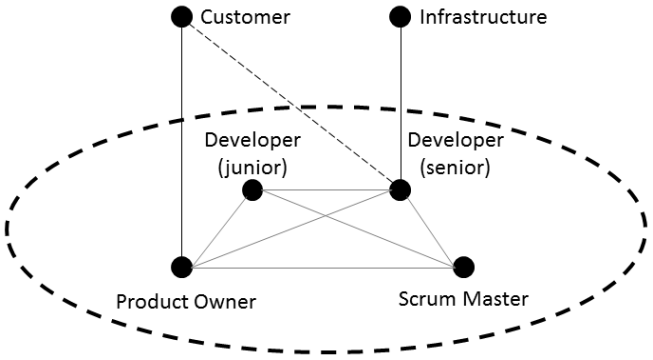


Fig. 5. Internal & external communication for Sunflower's Scrum team

A weekly meeting is scheduled with a representative of Infrastructure to prevent the Development Team from interfering with (scheduled) maintenance. However, whenever necessary, issues may also be taken up with Infrastructure immediately, not awaiting this meeting.

C. Local

The Scrum team of Logic operates in a larger organizational context: The Software Development department, consisting of:

- A Product group that is in charge of the implementation of software with new customers. Its developers convert customers from their previous software to Local's software; its consultants support customers with the set-up of the software and participate in courses for customers.
- Whenever a new customer approves of his software, it is taken into production at the customer and the responsibility within Local is transferred to the Support group. Support is the first point of contact for customers and functions primarily as a helpdesk. Support has some consultants for customer support on site or, again, courses, but does not employ developers. Changes as a result of customer reports are transferred to Development.
- Development deals with all modifications to the software, whether new features or as a result of bug reports. The Scrum team is found within this group.

Analogously to the Controller team Local's Scrum team uses all of the regularly scheduled Scrum 'meetings'. The figure below (Fig. 6) indicates communication within and beyond the team's borders.

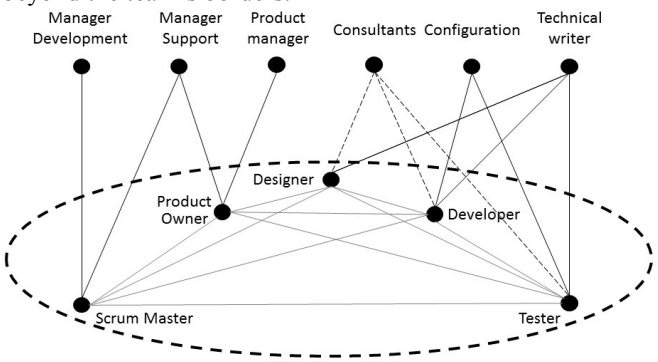


Fig. 6. Internal & external communication for Local's Scrum team

The managers (Development and Support) and the Product manager were already identified as stakeholders for customers. This implies that there is no direct communication between team members and partners outside of Local.

Configuration is in charge of the technical infrastructure. This partner, for instance, transfers software from a development stage to a testing stage to production (or vice versa). Involvement takes place in every Sprint.

The role of the technical writer is equal to the role of the documentarist in the Controller team.

D. Integrating results

We have shown results for the three individual teams with regard to their communication. In integrating these results, and in line with our research questions, we now establish a common vocabulary by mapping the partners from the individual cases to a limited set, enumerate them and indicate whether the communication is structural or ad hoc (Table 2).

V. Discussion

It is important to note that, for none of the teams, there were differences of opinion on their composition. All Scrum teams were crystal clear about their membership. As an example, when interviewing members of Local's Scrum team all members agreed on having ten members on the team and every team member mentioned the same persons in

TABLE II. SUMMARY OF RESULTS

External partner	**Controller**	**Sunflower**	**Local**
Management	*Management team*	-	*Manager Support*
			Manager Development
Consultant	*Consultants*	-	*Product manager*
			Consultants
Customer	*Customer*	*Customer*	-
Developer	*Frame/Web*	*Developer*	-
	Kanban team		
Documentarist	*Documentation*	-	*Technical writer*
Database specialist	*Database*	-	-
Infrastructure specialist	-	*Infrastructure*	*Configuration*

Structural communication
Ad hoc communication

the same roles (Table 1). Whether being involved full time or part time (some Scrum Masters and Product Owners), team membership, and thus also non membership, was incontrovertibly.

In the staffing of the teams it is straightforward that competencies of team members are in one or more of the 'traditional' phases of a software development life cycle: Analysis/design, programming, testing (Table 1, Fig. 3, 5 & 6); these are indeed engineering skills [17-18].

When looking at the partners outside teams' boundaries we distinguish three categories:

1. In a first category we unite those partners who, from the viewpoint of socio-technical congruence, match coordination requirements with actual coordination activities; in fact they are the partners to be expected, as already identified in Fig. 1.
2. In a second category we mention partners who do not match coordination requirements with actual coordination activities from the viewpoint of socio-technical congruence, but join the team through ad hoc communication.
3. The third category concerns partners who also do not match coordination requirements with actual coordination activities, but do so through structural communication.

This first category encompasses consultants, customers and management. None of these partners provides engineering skills. Instead they provide domain knowledge and their communication takes place via the Product Owner (with one small exception where the Scrum Master was also involved). This comes as no surprise, since the Product Owner is the linking pin with stakeholders as he is responsible for managing the Product Backlog, an ordered list of everything that might be needed in the product [4]. Management also communicates on (project) progress with the Product Owner and/or the Scrum Master, because in Scrum there is no role of project manager. In fact none is needed; the traditional responsibilities of a project manager have been divided up and reassigned among the three Scrum roles [29], especially the Product Owner and the Scrum Master [30]. Communication with management is thus to be expected; Scrum teams do not operate in a vacuum. In a previous study we already found that Scrum teams use project plans and progress information for control purposes [31]. And our current results show the structural character of this communication.

The second category consists of developers and database specialists, contributing engineering skills or competencies with regard to code/coding and database (structure) respectively. Often this communication is facilitated by the Scrum Master or arranged by the Product Owner; initiatives originate from these two roles, but often delegated to a developer working on a particular backlog item thereafter. The communication then is from one developer to another. This category is in fact no different from non-Scrum software development: In e-mail discussions in software-engineering communication emergent people were included in a discussion as a result of an explicit request [32].

From the viewpoint of socio-technical congruence our data show a clear mismatch between coordination requirements and actual coordination in the third category. Both documentarist and infrastructure specialist are structurally involved in the Scrum teams' activities. The documentarist is competent with regard to the production of user-related materials, such as a user guide. Communication is supported by data in a registration tool, such as (elaborated) user stories, design documents, test reports. The infrastructure specialist supports the Scrum team with regard to transition of code to a test or production facility; of course this is a part time activity; he also supports other teams. The incorporation of such a specialist in a Scrum team's activities coincides with the DevOps concept; this is an agile operations concept that uses agile techniques to link up departments - Development (Dev) and Operations (Ops) - together, which traditionally operated in silos [33].

No matter to which category partners belong, Scrum teams are well aware of their external partners and their competencies. Differences between the categories refer to whether or not the use of competencies was planned and whether it was structural or ad hoc. Revisiting the five

factors for coordinating expertise outside agile teams—availability, agile mind-set, stability, knowledge retention, and effective communication—these factors are in fact already incorporated in the Scrum teams' communication with their external partners.

In fact we introduce the notion of an extended Scrum team, which includes partners structurally involved in a team's activities, because communication occurs Sprint after Sprint (Fig. 7).

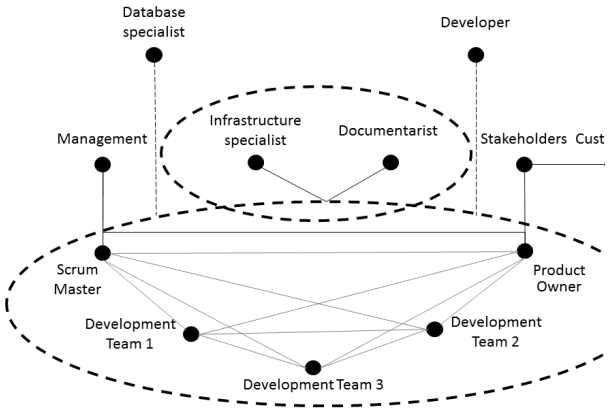


Fig. 7. The extended Scrum team

The extended Scrum team now incorporates an infrastructure specialist and a documentarist, adding technical and writing competencies. They work part time in the Scrum team, but they do not attend Scrum events, or at least not as a habit. The database specialist and the developer are (still) external partners. In other contexts it is conceivable that the database specialist is also a member of the extended Scrum team, because of his additional competencies.

III. Conclusions

Scrum team members agree on the boundaries of their team and are indeed cross-functional as far as 'traditional' phases of software development are concerned. Analysis/design, programming and testing competencies are represented in the teams.

To whom then do Scrum teams turn for additional competencies, which competencies are involved and how are Scrum teams aware of additional competencies they need?

- When additional competencies are required with regard to engineering skills, already represented in the team, the team turns to partners within the own organization. The teams are already aware of their partners and consult them on an ad hoc basis. They include other developers and database specialists.
- Additional competencies are also found in partners who structurally contribute their competencies to the Scrum teams in every Sprint. These partners include management, documentarists and infrastructure specialists. Here management was already expected to do so with regard to domain expertise and progress information; this is indeed a socio-technical match between coordination requirements and actual coordination activities. Others were not; especially documentarists and infrastructure specialists structurally contributed, although not full time, to the Scrum teams; they could be considered to be members of an extended Scrum team.

A. Limitations

In our three case studies documentarists and infrastructure specialists are found to be members of an extended Scrum team; a database specialist was not. This particular result cannot be generalized to Scrum teams in general; depending on, for instance, domain area, software type, or specific features of a team, partners could be distributed in another way, with the infrastructure specialist outside a team's border and, for instance, a database specialist inside. This argument also applies to partners who were not (even) identified in our case studies. What is generalizable, though, is the notion of the extended Scrum team, including partners not considered to be member of the team, but still structurally contributing to a team's activities. This allows practitioners to staff a Scrum team without pursuing overall cross-functionality in the team itself.

B. Future work

Of course we would like to see our results confirmed with other organizations. We believe the extended Scrum team to be an important extension of the notion of a cross-functional Scrum team and we would like to observe more and/or other external partners to elaborate this notion. We are also eager to pursue situational factors, whether organizational or personal, that determine the inclusion of external competencies.

Acknowledgement

The results from this research would not have been possible without the generous cooperation from the three case study organizations. The first author also wants to express his gratitude to Avans University of Applied Sciences for facilitating and supporting this research.

References

[1] D. Bustard, G. Wilkie, D. Greer, "The diffusion of agile software development: Insights from a regional survey", in Information Systems Development: Reflections, Challenges and New Directions, R. Pooley, J. Coady, C. Schneider, H. Linger, C. Barry, M. Lang, Eds., New York: Springer, 2013, pp. 219–230, http://dx.doi.org/10.1007/978-1-4614-4951-5_18.

[2] P. Rodríguez, J. Markkula, M. Oivo, K. Turula, "Survey on agile and lean usage in Finnish software industry", in Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '12), New York, 2012, pp. 139–148, http://dx.doi.org/10.1145/2372251.2372275.

[3] T. Chow, D.-B. Cao, "A survey study of critical success factors in agile software projects", Journal of Systems and Software, vol. 81, no. 6, pp. 961–971, June 2008, http://dx.doi.org/10.1016/j.jss.2007.08.020.

[4] K. Schwaber, J. Sutherland, "The Scrum guide – The definitive guide to Scrum: The rules of the game", 2013, retrieved from https://www.scrum.org/Portals/0/Documents/Scrum Guides/2013/Scrum-Guide.pdf.

[5] K. Schwaber, M. Beedle, "Agile software development with Scrum", 1st ed., Upper Saddle River (NJ): Prentice Hall, 2002.
[6] H.F. Cervone, "Understanding agile project management methods using Scrum", OCLC Systems & Services: International Digital Library Perspectives, vol. 27, no. 1, pp.18–22, 2011, http://dx.doi.org/10.1108/10650751111106528.
[7] K.B. Hass, "The Blending of traditional and agile project management", PM World Today, vol. IX, no. V, pp. 1–8, May 2007.
[8] D. Damian, S. Marczak, I. Kwan, "Collaboration patterns and the impact of distance on awareness in requirements-centred social networks", in Proceedings of the 15th IEEE International Requirements Engineering Conference (RE 2007), Delhi, 2007, pp. 59–68, http://dx.doi.org/10.1109/RE.2007.51.
[9] M. Mortensen, P. Hinds, "Fuzzy teams: Boundary disagreement in distributed and collocated teams", in Distributed Work, P. J. Hinds, S. Kiesler, Eds., Cambridge/London: MIT Press, 2010, pp. 283–308.
[10] K. Ehrlich, K. Chang, "Leveraging expertise in global software teams: Going outside boundaries", in Proceedings of the International Conference on Global Software Engineering (ICGSE '06), Florianopolis, 2006, pp. 149–158, http://dx.doi.org/10.1109/ICGSE.2006.261228.
[11] T. Chau, F. Maurer, "Knowledge sharing in agile software teams", in Logic versus Approximation - Essays Dedicated to Michael M. Richter on the Occasion of his 65th Birthday, W. Lenski, Ed., Berlin Heidelberg:Springer, 2004, pp. 173-183, http://dx.doi.org/10.1007/978-3-540-25967-1_12.
[12] J. Ferreira, H. Sharp, H. Robinson, "User experience design and agile development: managing cooperation through articulation work", Software: Practice and Experience, vol. 41, no. 9, pp. 963–974, August 2011, http://dx.doi.org/10.1002/spe.1012.
[13] A. Martin, R. Biddle, J. Noble, "An ideal customer: A grounded theory of requirements elicitation, communication and acceptance on agile projects", in Agile Software Development: Current Research and Future Directions, T. Dingsøyr, T. Dybå, N. B. Moe, Eds., Berlin Heidelberg: Springer, 2010, pp. 111–141, http://dx.doi.org/10.1007/978-3-642-12575-1_6.
[14] H. Sharp, H. Robinson, "Three "C"s of agile practice: Collaboration, Co-ordination and Communication", in Agile Software Development: Current Research and Future Directions, T. Dingsøyr, T. Dybå, N. B. Moe, Eds., Berlin Heidelberg: Springer, 2010, pp. 61–85, http://dx.doi.org/10.1007/978-3-642-12575-1_4.
[15] M. Cataldo, J.D. Herbsleb, K.M. Carley, "Socio-technical congruence: a framework for assessing the impact of technical and work dependencies on software development productivity", in Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement (ESEM 2008), Kaiserslautern, 2008, pp. 2–11, http://dx.doi.org/10.1145/1414004.1414008.
[16] I. Kwan, A. Schröter, D. Damian, "Does socio-technical congruence have an effect on software build success? A study of coordination in a software project", IEEE Transactions on Software Engineering, vol. 37, no. 3, pp. 307–324, May-June 2011, http://dx.doi.org/10.1109/TSE.2011.29.
[17] M. Kropp, A. Meier, "Teaching agile software development at university level: Values, management, and craftsmanship", in Proceedings of the IEEE 26th Conference on Software Engineering Education and Training (CSEE&T), San Francisco (CA), 2013, pp. 179–188, http://dx.doi.org/10.1109/CSEET.2013.6595249.
[18] P. Bootla, O. Rojanapornpun, P. Mongkolnam, "Necessary skills and attitudes for development team members in Scrum: Thai experts' and practitioners's perspectives", in Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), Songkhla, 2015, pp. 184–189, http://dx.doi.org/10.1109/JCSSE.2015.7219793.
[19] S. Downey, J. Sutherland, "Scrum metrics for hyperproductive teams: how they fly like fighter aircraft", in Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS), Wailea (HI), 2013, pp. 4870–4878, http://dx.doi.org/10.1109/HICSS.2013.471.
[20] M. Paasivaara, C. Lassenius, "Communities of practice in a large distributed agile software development organization - Case Ericsson", Information and Software Technology, vol. 56, no. 12, pp. 1556–1577, December 2014, http://dx.doi.org/10.1016/j.infsof.2014.06.008.
[21] J. Ferreira, H. Sharp, H. Robinson, "Values and assumptions shaping agile development and user experience design in practice", in Agile Processes in Software Engineering and Extreme Programming - Proceedings of the 11th International Conference on Agile Software Development (XP 2010), A. Martin, X. Wang, & E. Whitworth, Eds., Berlin Heidelberg: Springer, 2010, pp. 178–183, http://dx.doi.org/10.1007/978-3-642-13054-0_15.
[22] M.M. Rejab, J. Noble, G. Allan, "Distributing expertise in agile software development projects", in Proceedings of the Agile Conference (AGILE '14), Kissimmee (FL), 2014, pp. 33–36, http://dx.doi.org/10.1109/AGILE.2014.16.
[23] M.M. Rejab, J. Noble, G. Allan, "Locating expertise in agile software development projects", in Agile Processes in Software Engineering and Extreme Programming - Proceedings of the 15th International Conference (XP 2014), G. Cantone & M. Marchesi, Eds., Cham/ Heidelberg/ New York/ Dordrecht/ London: Springer International Publishing, 2014, pp. 260-268, http://dx.doi.org/10.1007/978-3-319-06862-6_19.
[24] M.M. Rejab, J. Noble, S. Marshall, "Coordinating expertise outside agile teams", in Agile Processes in Software Engineering and Extreme Programming - Proceedings of the 16th International Conference XP 2015, C. Lassenius, T. Dingsøyr, M. Paasivaara, Eds., Cham/ Heidelberg/New York/Dordrecht/London: Springer International Publishing, pp. 141-153, http://dx.doi.org/10.1007/978-3-319-18612-2_12.
[25] R.K. Yin, Case Study Research: Design and Methods (Applied Social Research Methods Series - volume 5), 4th ed., vol. 34, Thousand Oaks (CA): Sage Publications, Inc., 2009,
[26] H. Maimbo, "Designing a case study protocol for application in IS research", in Proceedings of the 9th Asia Conference on Information Systems (PACIS 2005), Bangkok, 2005, pp. 1281–1292.
[27] M. Paasivaara, C. Lassenius, "Collaboration practices in global inter-organizational software development projects", Software Process: Improvement and Practice, vol. 8, no. 4, pp. 183–199, October/December 2003, http://dx.doi.org/10.1002/spip.187.
[28] P. Runeson, M. Höst, "Guidelines for conducting and reporting case study research in software engineering", Empirical Software Engineering, vol. 14, no. 2, pp. 131–164, April 2009, http://dx.doi.org/10.1007/s10664-008-9102-8.
[29] P. Deemer, G. Benefield, C. Larman, B. Vodde, "The Scrum primer - A lightweight guide to the theory and practice of Scrum (Version 2.0), 2010, retrieved from http://www.goodagile.com/scrumprimer/scrumprimer20.pdf.
[30] M. Yilmaz, R.V. O'Connor, P. Clarke, "A systematic approach to the comparison of roles in the software development processes", in Software Process Improvement and Capability Determination - Proceedings of the 12th International Conference on Process Improvement and Capability dEtermination in Software, Systems Engineering and Service Management (SPICE 2012), A. Mas, A. Mesquida, T. Rout, R. V. O'Connor, A. Dorling, Eds., Berlin/Heidelberg: Springer, 2012, pp. 198–209, http://dx.doi.org/10.1007/978-3-642-30439-2_18.
[31] G. Wagenaar, R. Helms, D. Damian, S. Brinkkemper, "Artefacts in agile software development", in Product-Focused Software Process Improvement - Proceedings of the 16th International Conference on Product-Focused Software Process Improvement (PROFES 2015), P. Abrahamsson, L. Corral, M. Oivo, B. Russo, Eds., Springer International Publishing, pp. 133–148, http://dx.doi.org/10.1007/978-3-319-26844-6_10.
[32] I. Kwan,, D. Damian, "The hidden experts in software-engineering communication", in Proceedings of the 33rd International Conference on Software Engineering (ICSE '11), NIER Track, Waikiki, Honolulu (HI), 2011, pp. 800–803, http://dx.doi.org/10.1145/1985793.1985906.
[33] S. Karunakaran, "Impact of cloud adoption on agile software development", in Software Engineering Frameworks for the Cloud Computing Paradigm, Z. Mahmood, S. Saeed, Eds., London: Springer, 2013, pp. 213–234, http://dx.doi.org/10.1007/978-1-4471-5031-2_10.

Author Index