6^{th} International Conference on Information Technologies and Information Society
ITIS 2014

Proceedings

Edited by Zoran Levnajić and
Biljana Mileva Boshkoska
Faculty of Information Studies in Novo mesto

Šmarješke toplice, Slovenia, 5-7 November 2014
http://itis2014.fis.unm.si/

CONFERENCE ORGANIZATION

ORGANIZING COMMITTEE

Zoran Levnajić (chair), Biljana Mileva Boshkoska, Barbara Pavlakovič
Marjeta Grahek (administrative support)
Maja Zorčič (financial matters)

PROGRAM COMMITTEE (Paper referees)

Sanda Martinčić-Ipšić, University of Rijeka, Croatia
Bernard Ženko, Ježef Stefan Institute, Ljubljana, Slovenia
Markus Schatten, Faculty of Organization and Informatics, Croatia
Matteo Marsili, Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
Franjo Cecelja, University of Surrey, UK
Tijana Milenković, University of Notre Dame, USA
Marija Mitrović, Institute of Physics Belgrade, Serbia
Igor Bernik, University of Maribor, Slovenia
Miroslav Bača, Faculty of Organization and Informatics, Croatia
Aneta Stefanovska, Lancaster University, UK
Lovro Šubelj, University of Ljubljana, Slovenia
Nataša Pržulj, Imperial College London, UK
Antonina Dattolo, University of Udine, Italy
Davide Rossi, University of Bologna, Italy
Ljupčo Todorovski, University of Ljubljana, Slovenia
Matjaž Jurič, University of Ljubljana, Slovenia
Ana Meštrović, University of Rijeka, Croatia
Boštjan Delak, ITAD, Ljubljana, Slovenia
Marko Bohanec, Jožef Stefan Institute, Slovenia
Marija Mitrović Institute of Physics, Serbia
Panče Panov, Jožef Stefan Institute, Slovenia
Janez Povh, Faculty of information studies, Slovenia
Jože Bučar, Faculty of information studies, Slovenia
Marko Bajec, University of Ljubljana, Slovenia
Jernej Agrež, Faculty of Information studies, Slovnia
Neli Blagus, University of Ljubljana, Slovenia
Olivera Grljević, University of Novi Sad, Serbia
Tadej Kanduč, Faculty of information studies, Slovenia
Borut Lužar, Faculty of information studies, Slovenia

CONFERENCE PROGRAM

Thursday, November 6
08:30 - 09:00 Registration
09:00 - 09:10 Opening
09:10 - 10:00 Keynote 1: Taha Yasseri: What we could read from Wikipedia apart from its articles? Conflicts, Power, Fame, and Money
10:00 - 10:20 Sanda Martinčić-Ipšić: An Overview of Language Networks: Case of Croatian
10:20 - 10:40 Ana Meštrović: Network Motifs Analysis of Croatian Literature
10:40 - 11:00 Slobodan Beliga: Node Selectivity as a Measure for Graph-Based Keyword Extraction in Croatian News
11:00 - 11:30 *Coffee break*
11:30 - 12:20 Keynote 2: Boris Podobnik: Network risk and forecasting power in phase-flipping dynamical networks
12:20 - 12:40 Ana Meštrović: Toward Network-based Keyword Extraction from Multitopic Web Documents
12:40 - 13:00 Domagoj Margan: Toward a Complex Networks Approach on Text Type Classification
13:00 - 13:20 Jože Bučar: Machine Learning in Classification of News Portal Articles
13:20 - 13:40 Andrej Dobrovoljc: An approach to predict malicious threats
13:40 - 14:30 *Lunch break*
14:30 - 15:20 Keynote 3: Nataša Pržulj: Mining network data
15:20 - 15:40 Dragana Miljković: Constructing biological models from domain knowledge and literature
15:40 - 16:00 Rok Piltaver: Comprehensibility of Classification Trees - Survey Design Validation
16:00 - 16:20 Uroš Mesojedec: Assessing the potentials of cloud-native applications
16:20 - 16:40 Biljana Mileva Boshkoska: Alternative way of evaluation of air pollution levels
16:40 - 17:10 *Coffee break*
17:10 - 17:30 Jože Bučar: Case Study: Web Clipping and Sentiment Analysis
17:30 - 17:50 Boštjan Delak: Information System Mirror - Approach How to Analyze Information System Strengths and Weaknesses within Organization 17:50 - 18:10 Tomaž Aljaž: Bottleneck is on the top of the bottle or how to improve throughput in IT development
18:10 - 18:30 Albert Zorko: Mental disorder diagnostics from analysis of biomedical data - a review
18:30 - 19:30 *free*
19:30 - *Conference dinner and wine tasting*

Friday, November 7
10:00 - 10:50 Keynote 4: Franjo Cecelja: Ontology engineering: support to industrial processes
10:50 - 11:10 Jernej Agrež: A principled approach to the optimization solution of the biometric system
11:10 - 11:30 Darko Zelenika: Automatic invoice capture in small and medium-sized Slovenian enterprises - final report
11:30 - 11:50 Tadej Kanduč: Project of manufacturing processes optimisation in Podgorje Ltd.
11:50 - 12:10 Daniel K. Rudolf: Modern IT solutions in the logistics process
12:10 - 12:40 *Coffee break*
12:40 - 13:00 Andrej Dobrovoljc: Information security culture of online banking users
13:00 - 13:20 Valter Popeškić: Software defined network overview and security feature proposal
13:20 - 13:40 Petar Jurić: Game-based Learning and Social Media API in Higher Education
13:40 - 14:00 Andrej Kovačič: Consequences of importing negative news from abroad
14:00 - 15:00 *Lunch break*
15:00 - 15:20 Ljupčo Todorovski: Inferring Structure of Complex Dynamical Systems with Equation Discovery
15:20 - 15:40 Jaka Kranjc: Modeling wireless networks using graph theory
15:40 - 16:00 Jelena Govorčin: Extremal graphs with respect to vertex betweenness for certain graph families
16:00 - 16:20 Vesna Andova: Diameter on some classes of fullerene graphs
16:20 - 16:40 Stefana Janicijević: Comparison of SAS/STAT procedure and Variable Neighbourhood Search based clustering applied on Telecom Serbia data
16:40 - *Closing*

(only the name of the person presenting the contribution is provided, regardless of the total number of authors)

CONFERENCE SPONSORS

"Operacijo delno financira Evropska unija in sicer iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa krepitve regionalnih razvojnih potencialov za obdobje 2007-2013, 1. razvojne prioritete: Konkurenčnost podjetij in raziskovalna odličnost, prednostne usmeritve 1.1: Izboljšanje konkurenčnih sposobnosti podjetij in raziskovalna odličnost."

Table of Contents

Keynote speaker 1

What we could read from Wikipedia apart from its articles? Conflicts, Power, Fame, and Money

Taha Yasseri
Oxford Internet Institute, University of Oxford, UK

Abstract Wikipedia is the largest encyclopaedia in the world and seeks to "create a summary of all human knowledge". An encyclopaedia is supposed to contain a collection of objective facts, reported by secondary sources. However, the crowdsourced nature of Wikipedia makes it a source of information by itself reflecting the interests, preferences, opinions, and priorities of the members of its community of editors. By analysing the editorial conflicts between the editors of different language edition, we can create interesting images of each language community interests and concerns. Moreover, the page view statistics of Wikipedia articles, provide a unique insight to the patterns of information seeking by its readers. In this presentation, we start by Wikipedia edit wars and discuss what we could learn from the warring patterns about our real life facts, and then three examples are shown, in each of them statistics of editorial activities and page views are considered as proxies to assess popularity and visibility of items. Movie market, election, and scientific reputation are the three topics we have investigated and observed under certain conditions, there is a high correlation between popularity and Wikipedia edits and page views volumes. Based on these correlations and in the presence of external data to calibrate a predictive model, one is able to forecast the prospective success of an item in a reasonably accurate way.

Keynote speaker 2

Network risk and forecasting power in phase-flipping dynamical networks

Boris Podobnik
Faculty of Civil Engineering, University of Rijeka, Croatia

Abstract To model volatile real-world network behavior, we analyze a phase-flipping dynamical scale-free network in which nodes and links fail and recover. We investigate how stochasticity in a parameter governing the recovery process affects phase-flipping dynamics. We derive conditional probabilities for phase-flipping in networks. We apply our model to economic and traffic network.

Keynote speaker 3

Mining network data

Nataša Pržulj
Department of Computing, Imperial College London, UK

Abstract We are faced with a flood of network data. For example, various biomolecules interact in a cell to perform biological function, forming large networks. The challenge is how to mine these networks to answer fundamental questions about ageing, diseases and therapeutics. However, dealing with network data is non-trivial, since many methods for analysing large networks fall into the category of computationally intractable problems. We develop methods for extracting new biological knowledge from the wiring patterns of large molecular network data, linking network wiring with biological function and gaining new insight into diseases. We apply our methods to economics, tracking the dynamics of the world trade network that points to the origins of wealth and economic crises.

Keynote speaker 4

Ontology engineering: support to industrial processes

Franjo Cecelja
Process and Information Systems Engineering Research Centre, University of Surrey, UK

Abstract Industrial processes today are characterised by high dynamic and flexibility to adapt to new products and consequently by a short life time. Adding to it the unavoidable environmental component, Industrial Symbiosis (IS) is perhaps the most representative forms of integrated industrial processes. More precisely, IS is an innovative approach that aims in creating sustainable industrial networks set to process waste into materials, energy and water. Economic benefits are generated by normally low costs of waste or by-products, by using alternative energy sources and by environmental savings. Environmental benefits are inherent in IS and measured by landfill diversion but also by reduction in emissions and by water savings. Operating within confined geographic and administrative boundaries, IS also generates tangible social benefits to local communities, including job generation and retention, as well as new investments. The key to formation of IS networks is the mediation between participants, the process which requires expertise, hence knowledge in different areas, i.e. waste composition, capability of processing technologies and environmental effect, among the others. The whole process is currently managed by trained practitioners supported by proprietary databases. Limited to the level of expertise and intuition of practitioners and lacking readily available repository of tacit knowledge, the all operation is backward looking and focusing on past successful examples with innovative networks being incidental. This paper presents design and implementation of a semantic web platform which supports creation and operation of IS networks i) by screening the opportunities based on technological capability and resource availability of registered companies, and ii) by monitoring the IS operation and assessing sustainability using economic, environmental and social parameters. The platform employs ontologies to embed tacit knowledge in the domain of IS, knowledge gained from past experience but also from the latest research and otherwise advances in IS. More specifically, a set of integrated ontologies address off-spec nature of waste, i.e. variability in composition, dynamics in availability and pricing, as well as economic and environmental properties including hazardousness. Similarly, processing technologies are modeled in terms of processing capabilities, which include range of type of inputs, conversion rates, water and energy requirements, range of capacities, emissions as well as fixed and operational costs and environmental effects. Explicit knowledge is collected in the process of ontology instantiation with actual data collected from the IS participants during the registration. The ontologies are designed using ontology web language and hence prepared to grow and to share. In the current implementation more than 1500 different waste types and over 200 different technologies have been included. Purpose designed matchmaker identify synergies between participants on their semantic and explicit relevance, the process crucial to formation of IS networks. Semantic relevance defines suitability from the type of waste/by-product and range of technology inputs, including complex composites of waste, i.e. biodegradable waste, and it is calculated from distance between the two instances in the respective ontology. Semantic relevance also includes participant general suitability for particular type of IS. Explicit relevance is calculated using vector similarity algorithm for respective properties, such as quantity, availability, geographical location and hazardousness. More intuitive and complex IS networks are proposed by reclusively repeating matches between two participants which in turn gives an opportunity for even better economic and environmental savings and/or targeted production. Both semantic and explicit matching relevance are aggregated in into a numerical values use for match ranking. The platform has been implemented as a web service with performance validated verified in the industrial region in Viotia, Greece and with several hundred participating company. The effort has been funded by the LIFE+ initiative (LIFE 09 ENV/GR/000300), which authors acknowledge.

Overview of Language Networks: Case of Croatian

Sanda Martinčić-Ipšić, Ana Meštrović
Department of informatics
University of Rijeka
Croatia
{smarti, amestrovic}@uniri.hr

Abstract: *Viewed as a unique, biologically-based human faculty, language has been recognized as the reflection of the human cognitive capacities, both in terms of its structure and its computational characteristics. Studying languages at intra- and cross-linguistic levels is of paramount importance in relation to our biological, cultural, historical and social beings. Hence, human languages, besides still being our main tools of communication, reflect our history and culture. Language can be seen as a complex adaptive system, evolving in parallel with our society. The analysis and understanding of language and its complexity is of crucial interest in the modern globalizing world.*

Various aspects of natural language systems can be represented as complex networks, whose vertices depict linguistic units, while links/edges model their morphosyntactic, semantic and/or pragmatic interactions. Thus the language network can be constructed at various linguistic levels: syntactic, semantic, phonetic, syllabic ect. So far there have been efforts to model the phenomena of various language subsystems and examine their unique function through complex networks. Still, the present endeavors in linguistic network research focus on isolated linguistic subsystems lacking to explain (or even explore) the mechanism of their mutual interaction, interplay or inheritance. Obtaining such findings is critical for deepening our understanding of conceptual universalities in natural languages, especially to shed light on the cognitive representation of the language in the human brain.

Therefore, one of the main open questions in linguistic networks is explaining how different language subsystems mutually interact. The complexity of any natural language is contained in the interplay among several language levels. Below the word-level, it is possible to explore the type of phonology, morphology and syllabic subsystem complexity. The phonology subsystem complexity is reflected in the morphology subsystem complexity. On the word-level, the morphology subsystem complexity reflects in the complexity of the word order, syntactic rules and the ambiguity of lexis. Since the word order can be considered as the primary factor (but not the only) that determines linguistic structure, it is important to explore the subsystems' interactions by which it is influenced.
The overview of language networks will be revealed by means of the Croatian by showing the properties of five language subsystems: co-occurrence, syntax and shuffled layers (on the word level) and syllabic and grapheme layer (on the sub-word level).

Key Words*: complex networks, language…*

Network Motifs Analysis of Croatian Literature

Hana Rizvić, Sanda Martinčić-Ipšić, Ana Meštrović
Department of Informatics
University of Rijeka
Radmile Matejčić 2, Rijeka
{hrizvic, smarti, amestrovic}@uniri.hr

Abstract: *In this paper we analyse network motifs in the co-occurrence directed networks constructed from five different texts (four books and one portal) in the Croatian language. After preparing the data and network construction, we perform the network motif analysis. We analyse the motif frequencies and Z-scores in the five networks. We present the triad significance profile for five datasets. Furthermore, we compare our results with the existing results for the linguistic networks. Firstly, we show that the triad significance profile for the Croatian language is very similar with the other languages and all the networks belong to the same family of networks. However, there are certain differences between the Croatian language and other analysed languages. We conclude that this is due to the free word-order of the Croatian language.*

Key Words: *complex networks, linguistic networks, network motifs, triad significance profile, Croatian language*

1 Introduction

Many scientists from different disciplines study networks because of their ubiquity. The complex networks in nature share global properties such as small-world property of short paths between vertices and highly clustered connections [16]. In addition, many of these networks are scale-free networks, characterised by power-law degree distribution [2]. However, besides these global network characteristics, there are certain properties on the meso-scale and local-scale [4] that explain structural differences between complex networks. That is why more detailed network analysis on the meso-scale and on the local-level is important. Network analysis on the meso-scale and local-scale may include: community detection [12], motif analysis [10] or graphlet analysis [13].

In this paper we are focused on the network motifs' analysis. Network motifs are connected and directed subgraphs occurring in complex networks at numbers that are significantly higher than those in randomized networks [10]. Motifs may contain up to 8 vertices. For now, there have been reports on 3-vertex and 4-vertex motifs due to the complexity of the algorithm that identifies the motifs from the complex networks.

Alon et al. [11] analyse superfamilies of networks based on significant motifs (Fig. 1.). The first group of networks are from three microorganisms: the Escherichia coli, Bacillus subtilis and the Saccharomyces cerevisiae. These microorganisms form sensory transcription networks, the vertices represent genes or operons and the edges represent direct transcriptional regulation. They form the first superfamily which includes three types of biological networks: signal-transduction interactions in mammalian cells, developmental transcription networks arising from the review of the development of the fruit fly and sea-urchin, and synaptic wiring between neurons in Caenorhabditis elegans. They also studied three WWW networks of hyperlinks between web pages related to university, literature and music. A feature of these networks is the transitivity of the relations, as evidenced by the motifs presented in these networks that are highly transitive. Similar results are obtained by testing three social networks, where people from the group are represented by vertices. The connections between two people, a positive opinion of one member of the group to another member, were represented by edges, obtained on the basis of questionnaires. The conclusion is that social networks and the web are probably members of the same superfamily, which may facilitate the understanding of the structure of the web. Furthermore, word-adjacency networks are analysed so that each vertex represented a single word, and each edge represented a connection between the two words that have followed one another in the text. The results obtained for different texts in different languages (English, French, Spanish and Japanese) are similar. Significant triads are from ID_3#1 to ID_3#6 (considering the IDs in [11]), and underrepresented are all other triads, from the ID_3#7 to ID_3#13. This means that the examined languages do not have a transitive relation such as the WWW. The explanation for these results may be in the structure of language, where words are divided into categories and generally the rule is that a word from one category follows a word from the other category. As an example, most connected category words are prepositions and behind them usually follows a noun or an article.

Biemann at al. [3] use motifs to quantify the differences between a natural and a generated language. The frequencies of three-vertex and four-vertex motifs for six languages are compared with artificially generated language from *n*-grams. An *n*-gram is contiguous sequence of *n* units (words) reflecting the statistical properties of a given text (or speech). The authors show that the four-vertex motifs can be interpreted by semantic relations of polysemy and synonymy of words.

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

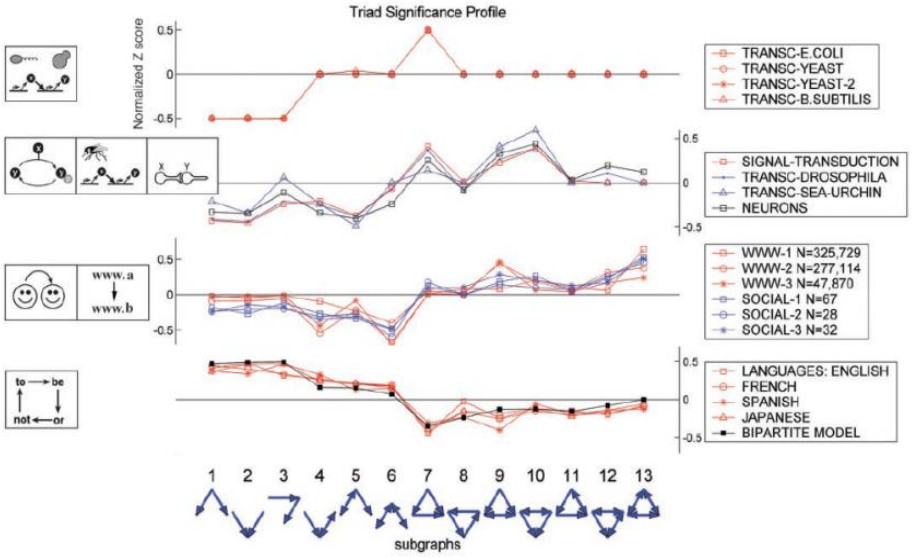


Figure 1: Superfamilies of complex networks according to the triad significance profile [11]

Our motivation for this research was to determine whether the local structure of the Croatian language networks share the same properties as other language networks. Croatian is a highly flective Slavic language and words can have seven different cases for singular and seven for plural, genders and numbers. The Croatian word order is mostly free, especially in non-formal writing. These features position Croatian among morphologically rich and free word-order languages. So far Croatian has been quantified in a complex networks framework based on the word co-occurrences [7], [1] and compared with shuffled counterparts [8], [9].

In this paper we describe the network motifs analysis of the co-occurrence directed networks constructed from the Croatian texts: four books and one forum. We use an approach based on the significance profile (*SP*) presented in [10]. We analyse three-vertex subgraphs called triads and present the results of triad significance profile (*TSP*) for the five analysed networks. In this paper we compare our results with *TSP* for other languages.

In the second section we give an overview of network motifs. In the third section we describe the experiment, and the fourth section presents the results. We conclude with some finishing remarks and the plans for future work.

2 Network motifs

A network motif is a small subgraph that appears more frequently in the real network than in the random network. The motif may be referred to as a significantly overrepresented subgraph in the network. As well, an underrepresented subgraph in the network is called an anti-motif. In [10] authors define network motifs as small patterns for which the probability of occurrence in a randomized network is less than the probability of occurrence in the real network with the cut-off value equal to 0.01.

In Fig. 2. are all 13 possible three-vertex connected directed subgraphs (triads). The triad ID notation in this paper is preserving the same notation as on the Fig 2 and it is the notation according to [11].

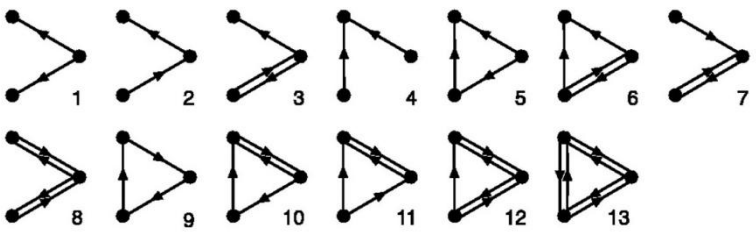


Figure 2: All 13 types of three-vertex connected subgraphs

In Fig. 3. are all 199 possible four-vertex connected directed subgraphs.

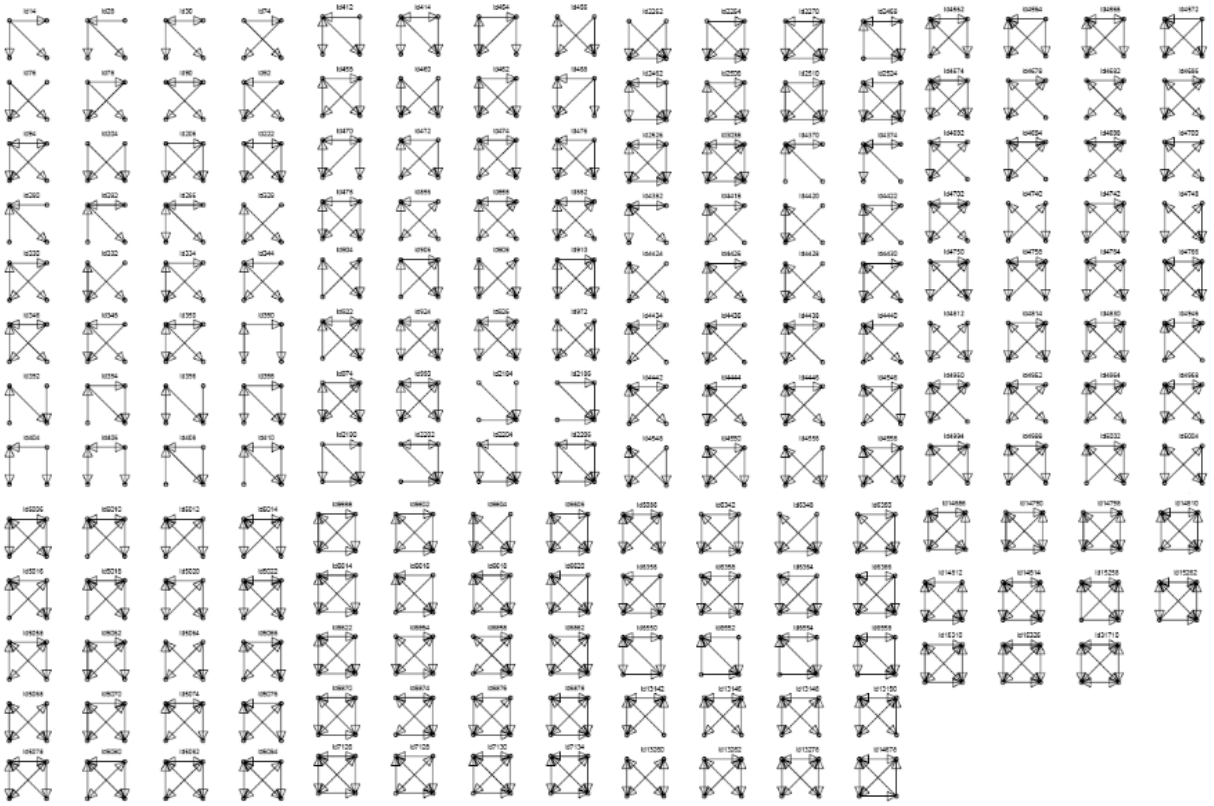


Figure 3: Four-vertex connected and directed subgraphs

Now, we will give the mathematical description of the motif in the graph or network *G*. There are two graphs (networks) *H* and *G* with non-empty sets of: vertices, edges and incidence relation. Let *H* be the real subgraph of *G*, $H \subset G$. The number of occurrences of graph *H* in graph *G*, we define as the frequency of *H* in *G*, written like $F_H(G)$. Some graph is frequent in *G* if its frequency in *G* is higher than cut-off value. Let $\Omega(G)$ be a family of randomized graphs of *G* (randomized graph has the same number of vertices and same degree sequence [10]). Now we choose *n* random graphs from $\Omega(G)$ uniformly, $R(G)$. Then we find out the frequency of the certain frequent sub-graph *H* in *G*. If the frequency of *H* in *G* is higher than its arithmetic mean frequency in *n* random graphs R_i, where $1 \leq i \leq n$, we call this sample significant and *H* is network motif for *G*.

Besides the frequency, motifs can be detected by using probabilities. The *p*-value of the motif is the number of random networks in which a particular motif appeared more frequently than in the original network, divided by the total number of generated random networks. Obviously, the *p*-value is between 0 and 1. The smaller the *p*-value of the motif is, the more significant the motif is.

Another measure for motif detection is a *Z*-score. The *Z*-score for the subgraph *H* in *G* can be calculated using the equation:

$$Z(H) = \frac{F_G(H) - \mu_R(H)}{\sigma_R(H)} \tag{1}$$

where $\mu_R(H)$ is the mean and $\sigma_R(H)$ is the standard deviation of frequencies of *H* in the set of random graphs of *G*, $R(G)$. The higher the *Z*-score is, the more significant a detected motif is. Using eq. 1, for each subgraph *i*, we can calculate the statistical significance which is described as *Z*-score, Z_i.

Furthermore, the *SP* is the vector of *Z*-scores normalised to length 1:

$$SP_i = \frac{Z_i}{\sqrt{\sum Z_i^2}} \tag{2}$$

3 Experiment

3.1 Datasets and networks construction

In our study, we examined five literary works. Our dataset contains five different texts; four books: *Mama Leone* (*ML*), *The Return of Philip Latinowicz* (*PL*), *The Picture of Dorian Gray*, (*DG*), *Bones*, (*BO*) and one forum *Narodne novine* (*NN*). All the books were written in or have been translated into Croatian. The web forum is selected as a representative of a different text genre in order to verify whether the observed properties are also valid for more relaxed genres besides those strictly for the literature.

The datasets are different in the size as well as in the size of the vocabulary (Table 1).

The texts were cleared of the index of contents, the authors' biographies and page numbers. Then we constructed directed co-occurrence networks (word-adjacency networks) in a way that each word represents a vertex, and the two words that follow one another establish the edge.

Table 1: Number of words, vertices and edges in the analysed datasets

Dataset	Number of words	Number of vertices (*N*)	Number of edges (*K*)
ML	86,043	12,416	52,012
PL	28,301	9,166	22,344
DG	75,142	14,120	47,823
BO	199,188	25,020	106,999
NN	146,731	13,036	55,661

3.2 Network motifs analysis

To analyse the motifs in networks we used the FANMOD tool [16]. FANMOD can search for motifs of three to eight vertices sizes using the rand-esu algorithm [15], which is much faster than similar tools, and the advantage is that it has a simple graphical interface and it is very intuitive to use.

The first step is the preparation of the input data: conversion of words to integers, where every number represents one vertex uniquely in the network, hence two integers in a line form an edge. Every line must contain at least two integers and a maximum of up to five integers. FANMOD provides the possibility to choose whether the networks have directed, undirected or coloured edges or vertices. We used directed uncoloured networks.

The algorithm options frame must be adjusted prior to running the algorithm itself. The options' frame includes: the set of the subgraph size and the setting of the switch between full enumeration and enumeration on a few samples. Motifs are identified through the comparison of frequencies in the original network and those in a random network so it is important to determine the number of random networks. It can be set up in the random networks frame in the box named 'Number of networks'. The default value for this is 1,000 networks but it can be increased if necessary. In this frame there are some important parameters: the parameter "exchanges per edge" (showing how many times the program goes through the edges) should be increased only if our results (output after the first reading) for a random network are very similar to the results for the original network. The parameter "exchange attempts" - if in the results there appears a small number of successful replacements, then we need to increase it, but it is important to bear in mind that if we have a few successful replacements it may mean that something is wrong with the network.

FANMOD produces results in terms of: *Z*-scores, *p*-values and frequencies. When we analyse the results, it is desirable to obtain as much as possible undefined *Z*-scores. If we have a lot of undefined *Z*-scores, it is not possible to determine which motif is significant (because the greater the Z-score is, the greater significance of this motif is). So if we have a lot of undefined *Z*-scores we have to increase the number of random networks, which will slow down the algorithm.

In the output file format is advisable to include an ASCII – text option for the easier reading of the results, and in HTML format for the presentation of the results. We calculate *Z*-scores for all triads in all five networks using FANMOD. After that we calculate *TPS* according to the eq. 2.

4 Results

The frequencies of all possible triads for five networks are presented in Fig. 4. In general, the triad frequencies behave similarly for all five networks. Therefore the Croatian language is comparable with other languages [3], [11]. Still, it is possible to identify differences between data source on ID_3#1 and ID_3#5 on the linear scale and ID_3#9, ID_3#11 and ID_3#13 on the logarithmic scale.

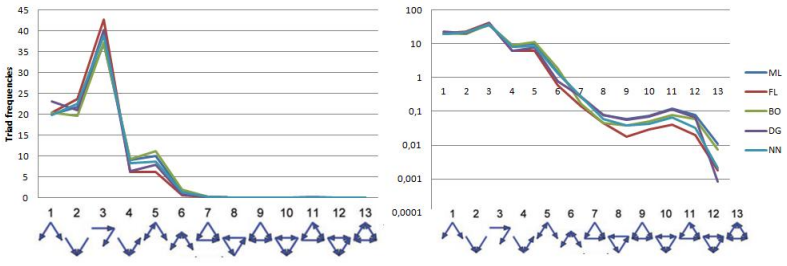


Figure 4: The frequencies of the triads for 5 datasets presented on the linear scale (left) and on the logarithmic scale (right)

Furthermore, we analyse *TSP* in order to detect which triads are significantly overrepresented (motifs) and which triads are significantly underrepresented (anti-motifs) and to compare it across the five different datasets. The results are presented in the diagram shown in the Fig. 5.

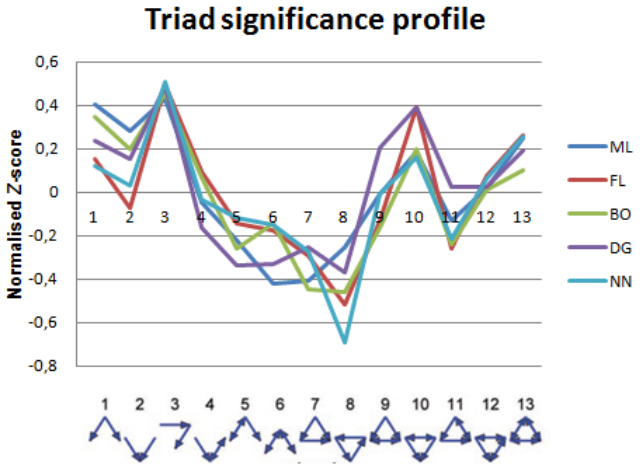


Figure 5: Triad significance profile for 5 datasets

There are several significantly overrepresented triads (ID_3#1, ID_3#3, ID_3#10 and ID_3#13). Triads with two edges (ID_3#1 and ID_3#3) are, based on the other reported results [3], [11], expected to be overrepresented in language networks. However, in our results, triads ID_3#10 and ID_3#13 are not likely to be overrepresented in language. It seems that this is inherent to languages with a free word-order such as Croatian. For example for three vertices of words: jako (*much*), ga (*him*), voli (*loves*); in Croatian language it is possible to have all six pairs of words (even triplets) as it is shown in Fig. 6. In opposite, in English language is impossible to have "him loves" as a part of the sentence.

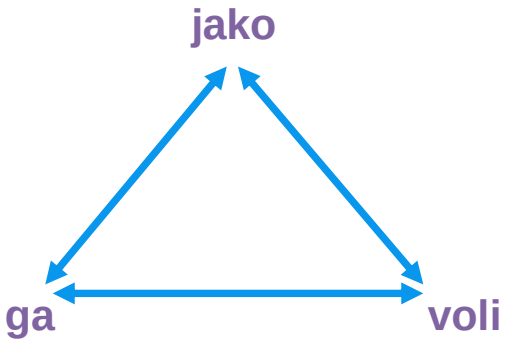


Figure 6: An example of the triad with ID_3#13 in Croatian language

5 Conclusion

In this paper we present the results of the network motifs analysis of Croatian literature. Motifs are used to detect structural similarities between directed networks of four books and one forum. We analyse triad significance profile in five different texts represented as directed co-occurrence networks.

The results show that Croatian language networks have similar triad significance profiles with other already analysed languages. Generally, in all language networks triads with two edges are overrepresented, while triads with three edges are underrepresented. For the Croatian language, there is an exception with three-edge triads ID3#10 and ID3#13 which are overrepresented. The overrepresentation of three-edge triads is caused by the free word-order nature of Croatian language.

It seems that motif-based analysis of the language networks is sensitive to the word order and syntax rules. And maybe it is possible to use it for the fine-grained differentiation of languages. Therefore, we will perform motif-based analysis of language networks for different languages. We will also include syntax networks and sub-word level networks (syllable networks, grapheme networks) in the analysis. Finally we plan to analyse the presence of the four-vertex motifs in language networks in order to see if they can be interpreted by the semantic relations in the polysemy and synonymy of words.

6 References

[1] Ban, K, Martinčić-Ipšić S., Meštrović, A. Initial comparison of linguistic networks measures for parallel texts. In 5th International Conference on Information Technologies and Information Society (ITIS), pp. 97-104. 2013.

[2] Barabási, A.- L., Albert, R. Emergence of scaling in random networks. Science 286, no. 5439 (1999): 509-512.
[3] Biemann, C., Roos S., Weihe, K. Quantifying semantics using complex network analysis. COLING12. 2012.
[4] Borge-Holthoefer, J., Arenas, A. Semantic networks: structure and dynamics. Entropy 12, no. 5 (2010): 1264-1302.
[5] Cancho, R. – F., Solé, R. The small world of human language. Proceedings of the Royal Society of London. Series B: Biological Sciences 268, no. 1482 (2001): 2261-2265.
[6] Hagberg, A., Swart, P., Chult, D. Exploring network structure, dynamics, and function using NetworkX. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Laboratory (LANL), 2008.
[7] Margan, D., Martincic-Ipšic, S., Meštrovic, A. Preliminary report on the structure of Croatian linguistic co-occurrence networks. 5th International Conference on Information Technologies and Information Society (ITIS), 89-96, 2013.
[8] Margan, D, Martincic-Ipšic, S., Meštrovic, A. Network Differences between Normal and Shuffled Texts: Case of Croatian. In Complex Networks V, pp. 275-283. Springer International Publishing, 2014.
[9] Margan, D., Meštrović, A., Martinčić-Ipšić, S. Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts. In IEEE 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2014). 2014
[10] Milo, R.; Shen-Orr, S.; Itzkovitz, S. et al. Network motifs: simple building blocks of complex networks. Science, 298(5594): 824 – 827, 2002
[11] Milo, R; Itzkovitz, S; Alon, U. et al. Superfamilies of evolved and designed networks. Science, 303(5663): 1538 – 1542, 2004.
[12] Newman, M. EJ. The structure and function of complex networks. SIAM review 45, no. 2 (2003): 167-256.
[13] Pržulj, N. Biological network comparison using graphlet degree distribution. Bioinformatics 23, no. 2 (2007): e177-e183.
[14] Rasche, F; Wernicke, S. FANMOD fast network motif detection – manual, Bioinformatics, 22(9):1152–1153, 2006.
[15] Watts, Duncan J., and Steven H. Strogatz. Collective dynamics of 'small-world'networks. nature 393, no. 6684 (1998): 440-442.
[16] Wernicke, S. A faster algorithm for detecting network motifs. In R. Cassadio and G. Myers, editors, Proceedings of WABI '05, number 3692 in LNBI, pages 165–177. Springer-Verlag, 2005.

Node Selectivity as a Measure for Graph-Based Keyword Extraction in Croatian News

Slobodan Beliga, Sanda Martinčić-Ipšić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{sbeliga, smarti}@inf.uniri.hr

Abstract: *In this paper, we introduce selectivity-based keyword extraction as a new unsupervised method for graph-based keyword extraction. Node selectivity measure is defined as the average weight distribution on the links of a single node and used in procedure of keyword candidate extraction. In particular, we propose extracting three word long keyword sequence and proving that the obtained results compare favourably with previously published results. Experiments were conducted on Croatian news articles dataset with keywords annotated by human experts. The selectivity-based keyword extraction method achieved the average F2 score of 25.32% on isolated documents and F2 score of 42.07% on a document collection. Proposed method is derived solely from statistical and structural information, which are reflected in the topological properties of text network. Furthermore, comparative results indicate that our simple graph-based method provides results that are comparable with more complex supervised and unsupervised methods, as well as with human annotators.*

Key Words*:* *keyword extraction, keyword candidate, keyword ranking, keyword expansion, node selectivity, Croatian news, complex network*

1 Introduction

Keywords are the terms that represent the most relevant information contained in a document. They represent specific metadata that can be useful to many IR (*Information Retrieval*) and NLP (*Natural Language Processing*) tasks: document indexing, document summarization, text/document/website categorization or clustering, etc. While manual assignment of keywords to documents is very costly and time consuming, the number of digitally available documents is growing and automatic keyword extraction attracts the researcher's interest. Although the keyword extraction applications usually work on a single document, keyword extraction can also be applied to the whole collection [1], the entire web site or the automatic web summarization [2].

Approaches to keyword extraction are supervised or unsupervised: supervised approaches require annotated data source [3-5, 26, 27], while unsupervised require no annotations in advance [1, 6, 9-15, 18, 20, 25]. One of the unsupervised methods is a graph enabled extraction in which the document is represented as a graph or a network of words connected in accordance with their relations in text (co-occurrence, syntax, semantic [12, 7, 17]). The network-based keyword extraction exploits different network

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

measures in order to identify and rank the most representative features of the source – the keywords from the text. Network properties are commonly measured by certain centrality measures such as degree, betweenness, closeness, strength, coreness, eigenvector centrality [11, 12] or by the modification of PageRank or HITS algorithm [11-13, 19, 25].

In this paper, we propose the Selectivity-Based Keyword Extraction (SBKE) method as a new unsupervised method for network-based keyword extraction. This work expands our initial idea of one and two word long keywords [8] to three word long sequences. Additionally, we show that the SBKE method is suitable for keyword extraction from the whole document collection of Croatian news articles (HINA dataset with manually annotated keywords) instead of keyword extraction from isolated documents. This SBKE method is designed as an architected solution for the ranking and extraction of keywords.

The paper is structured as follows: Sec. 2 describes related work on keyword extraction; Sec. 3 defines node selectivity measure; Sec. 4 elaborates SBKE architecture in two steps: keyword candidate extraction, and candidate expansion; Sec. 5 describes dataset, experiment, evaluation and obtained results with short discussion. In the last part the concluding remarks and appointments for future work are given.

2 Related Work

Lahiri et al. [11] extracted keywords and keyphrases from word co-occurrence networks and noun phrase collocation networks. Results on four data sets suggest that centrality measures outperform the baseline Term Frequency/Inverse Document Frequency (TF/IDF) model and simpler centrality measures like coreness and betweenness. Boudin reviewed various centrality measures for keyword extraction for English and French datasets [12]. His experiment shows that simple degree centrality achieves results comparable to the widely used TexRank algorithm. PageRank and its modification - TextRank were compared to centrality measures in Mihalcea experiments [13]. However, these measures were also compared with a supervised machine learning n-gram based approach. It has shown that the degree centrality gives similar results as TextRank, while closeness centrality outperforms TextRank on short documents. In addition, it is noted that the PageRank of the graph-based approach performs analogously as supervised machine learning approach. There are a number of other graph-based research reports (supervised or unsupervised) proposing keyword extraction for the task of extractive summarization (using HITS algorithm and degree-based ranking) [14], key terms extraction with community detection techniques [15], text summarization by extracting most important sentences (LexRank, the concept of eigenvector centrality) [16] or keyword and sentence extraction using SemanticRank (variations of PageRank and HITS) [17]. All of them appear to be promising graph-based extracting and ranking approaches. It is worth noticing that reported research relay on simple measures, like centrality (degree, closeness, …) and achieve surprisingly better results than more complex measures (eg. TextRank) [11, 12].

The keyphrase extraction for Croatian language includes supervised [3] and unsupervised approaches [9, 10, 18]. In a supervised approach keyphrase candidates are generated by linguistic and statistical features (relative first appearance in document, TF/IDF, etc.), and Naïve Bayes classifier is used to select the best keyphrases among candidates. Better results are obtained through unsupervised methods. In [9] part-of-

speech (POS) and morphosyntactic description (MSD) tag filtering followed by TF/IDF ranking are used. They brought up the conclusion that simple filtering provides results comparable to those of the human annotators. Saratlija et al. in [18] use distributional semantics to form topically related clusters, to extract keywords and to expand them to keyphrases. Finally, Bekavac and Šnajder in [10] propose genetically programmed keyphrase extraction method for Croatian language - GPKEX. They represent a keyphrase scoring measure as syntax trees and evolve them to produce rankings for keyphrase candidates. Obtained results are comparable to more complex machine learning methods previously developed for Croatian. All mentioned researches for Croatian keyphrase extraction used dataset composed of Croatian news articles from the Croatian News Agency (HINA), with manually annotated keywords by human experts.

3 Selectivity

Node selectivity was originally proposed by Masucci and Rodgers [22, 23] (a measure that can distinguish a regular and shuffled network [24]). It is actually the average strength of a node. In the network G, N is the number of nodes and K is the number of links [22]. In weighted language networks, every link connecting two nodes i and j has an associated weight w_{ij} that is a positive integer number. For the node i, the selectivity is calculated as a fraction of the node weight and the node degree:

$$e_i = \frac{s_i}{k_i} . \quad (1)$$

In other words, node selectivity is defined as the average weighted distribution on the links of a single node. In the directed network, the in/out selectivity of the node i is defined as:

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}} , \quad (2)$$

where $s_i^{in/out}$ in directed network is the in/out strength of the node i, and is defined as the number of its incoming and outcoming links:

$$s_i^{in/out} = \sum_j w_{ji/ij} . \quad (3)$$

In addition, the node degree k_i is defined as the number of edges incident upon a node. The in-degree and out-degree $k_i^{in/out}$ of the node i is defined as the number of its in and out neighbors.

4 Selectivity-Based Keyword Extraction

This paper is an extension of our previous research in which the selectivity of a node is presented as a measure for extracting the keywords from news articles [8]. In [8] we also showed that selectivity slightly outperforms the centrality-based approaches: in-degree, out-degree, betweenness and closeness. Nodes with the highest selectivity value are open-class words (content words) which are preferred keyword candidates (nouns, adjectives, verbs) or even part of collocations, keyphrases, names, etc. Selectivity is insensitive to non-content words or stop-words (the most frequent function words, which do not carry strong semantic properties, but are needed for the syntax of the language) and therefore can efficiently detect semantically rich open-class words from the network and extract keyword candidates. To the best of our knowledge, the node selectivity measure was not applied to keyword extraction task before. It was a motivation for continuing research on the trail of selectivity measure and improvements of the existing algorithm for better keywords expansion with relation to the node with high selectivity.

In this paper we extend the method to three words long sequences. Proposed Selectivity-Based Keyword Extraction (SBKE) is fully unsupervised method and consists of two phases: keyword candidate extraction (based on selectivity measure) and keyword expansion (keyword two-expansion – K2E; keyword three-expansion – K3E).

4.1 Keyword Candidate Extraction

Keyword candidate extraction can be divided into two steps: the construction of co-occurrence network and calculating selectivity values for extraction. Text is represented as a complex network of related words: each individual word is a node and links are interactions amongst words. Co-occurrence networks exploit simple neighbor relation: two words are linked if they are adjacent in a window of maximum *n* words. Neighbor relation is preserved within the sentence boundaries. Punctuation such as [.,!?] indicate the end of co-occurrence. Links are weighted according to the co-occurrence frequencies of the words they connect [28]. Since Croatian is a highly inflectional Slavic language, a source text usually needs a substantial preprocessing (lemmatization, morphological normalization, stop-words removal, part-of-speech (POS) annotation, morphosyntactic description (MSD) tagging, etc.). We have designed our approach with light or no linguistic knowledge. The preprocessing includes conversion of the input text to lowercase, cleaning of misspelled symbols (w instead vv, ! instead l, 0 instead O, etc.) and lemmatization. Lemmatization was conducted to circumvent the effects of morphological variations. Non-standard word forms like numbers, dates, acronyms, abbreviations etc., remain in the text, since the method is preferably resistant to the noise presented in the data source. It is important to note that in order to retain subtle distinctions in the text and actual word order from sentences, unlike similar graph-based approaches in [11, 12, 13 19] or other supervised and unsupervised approaches for Croatian presented in [3, 9, 10], words added to the network in our approach are not restricted to any syntactic filtering, like POS tagging - for selecting certain words (open class words: i.e. nouns or adjectives).

The second step determines the importance of each node from the network based on node selectivity value. When selectivity score $e_i^{in/out}$ is computed for all nodes in all 60 networks, the nodes are ranked according to the highest in/out selectivity value above threshold value. Preserving the same threshold value ($e_i^{in/out} \geq 1$) in all documents different numbers of nodes rise (one word long keyword candidates) extracted from each network. A set obtained from one word long keyword candidates is noted as SET1 and is retained for second phase: keyword expansion.

4.2 Keyword Expansion

Keyword expansion is tasked with the detection of neighboring nodes for filtered candidates with high selectivity value. Keywords are expanded to two word long sequences of keyword candidates (words-tuples) and to three word long sequences of keyword candidates (words-triples). The first part of the expansion is called the keyword-two expansion (K2E). For nodes that have passed in-selectivity filtering (e_i^{in}), we isolate one neighbor node with the highest outgoing weight ($max\ w_i^{out}$) - predecessor. The same procedure is applied to out-selectivity filtering (e_i^{out}), but here we isolate one neighbor node with the highest ingoing weight ($max\ w_i^{in}$) - successor. The results of in/out selectivity extraction are a set of ranked words-tuples, noted as SET2. Words-tuples are two word long sequences of keyword candidates.

The second part of the expansion is called the keyword-three expansion (K3E). The procedure is very similar to that for K2E. However, expansion predecessor for in-selectivity or expansion successor to the out-selectivity is carried out over all expanded words-tuples candidates from the initial expansion, but not over one word long keyword candidates. So, K2E is the basis for K3E. The expansion of any predecessor or successor is always guided by the node with the highest in/out weight value. The result of in/out selectivity extraction in K3E is a set of ranked words-triples, noted as SET3. Previously explained K2E and K3E procedure for the case of predecessors' (at the top) and successors' expansion (at the bottom) are in Fig. 1.

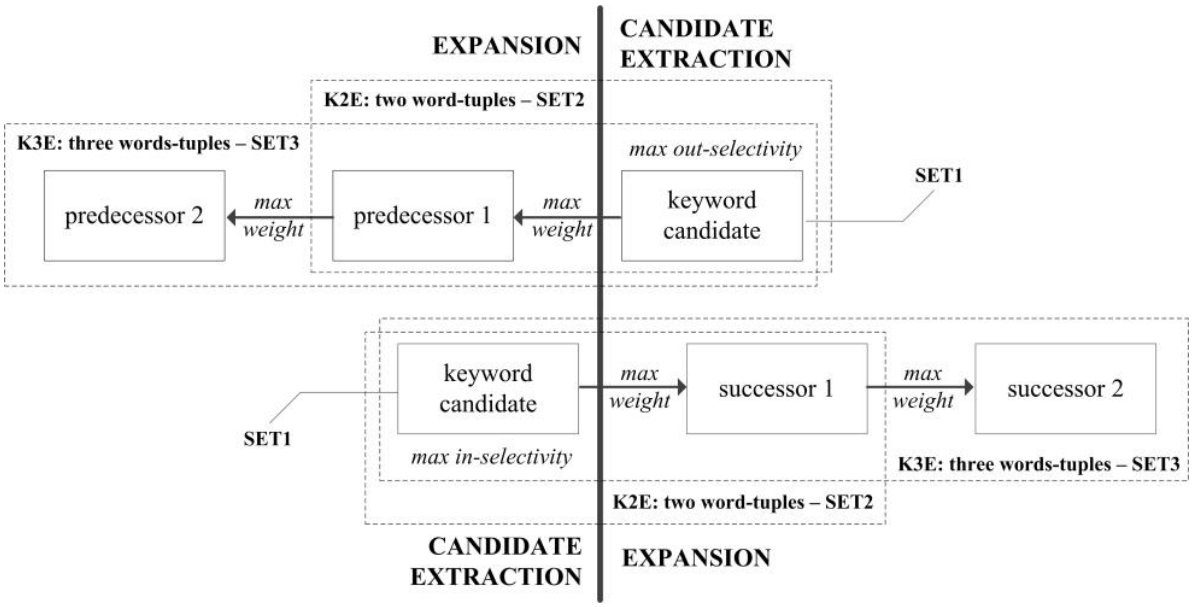


Figure 1: Two word (K2E) and three word (K3E) expansion

5 Evaluation and Results

5.1 Dataset

Although Selectivity-Based Keyword Extraction (SBKE) on co-occurrence word networks has been used before [8], there is no systematic or complete comparison with existing approaches (either supervised or unsupervised) on established benchmark datasets. The existing comparison is performed only on a partial dataset. In this work we have extended the evaluation to the whole dataset (60 documents) as explained in Sec. 1. The whole dataset contains of 1,020 news articles from Croatian News Agency (HINA). The keywords (keyphrases) in documents were annotated by eight human expert annotators. The set was divided into two parts: the first part - 960 annotated documents for supervised learning and the second part - 60 for testing. The subset of 60 documents was annotated by all eight annotators and used as "the golden set" for testing. The inter-annotator's agreements in terms of F2 scores were on average 46% (between 29.3% and 66.1%) [9]. The selected 60 test texts varied in the length from very short 60 up to 1500 tokens (335 on average) and contained 20,125 tokens in total. Number of annotated keywords per text varied between 9 and 42 (24 on average) and was nearly 1500 of all 60 texts. One annotator has annotated on average 10 keywords per document.

5.2 Keyword Extraction Experiment

From the HINA dataset we constructed 60 co-occurrence language networks, one for each text. According to the proposed SBKE method, we extracted keyword candidates and expanded them as explained in Sec. 4. In the same way we also constructed an integral network of all 60 news articles from the collection in order to gauge of how selectivity-based keyword extraction approaches larger and smaller networks, i.e. long and shorter texts. This poses the question whether SBKE method will achieve better results on a larger collection or on individual texts?

In both cases, obtained tuples and triples (Sec. 4.2) the stop-words are filtered out to enable a comparison with manually annotated test keywords. Stop-words are closed class words and can be part of a keyphrase, but they are certainly not standalone keywords. It should also be noted that the evaluation is valid if compared sets comply with the same criteria. For this reason, the stop-words are excluded from human annotated keywords, too. Thus, the evaluation was conducted between selectivity-based extracted keywords and human experts annotated keywords.

Network construction and analysis were implemented with the Python programming language using the NetworkX software package development for creation, manipulation and study of the structure, dynamics and functions of complex networks [21]. In the following section, detailed description of SBKE method is given.

5.3 Evaluation

Evaluation is the final part of the experiment based on the intersection of the obtained sets (SET1, SET2 and SET3) of keyword candidates with the union of all 8 human annotated keywords. We evaluate the performance of SBKE methods in terms of recall (*R*), precision (*P*) and *F1* score. Recall is calculated as the fraction of the human annotated keywords which is intersected with keywords extracted with a SBKE method and all human annotated keywords:

$$R = |\{human\ annotated\} \cap \{SBKE\}| / |\{\ human\ annotated\ \}|. \quad (4)$$

Precision is calculated as the fraction of the human annotated keywords which is intersected with keywords extracted with a SBKE method and SKBE keywords:

$$P = |\{human\ annotated\} \cap \{SBKE\}| / |\{SBKE\}|. \quad (5)$$

F1 score is the weighted harmonic mean of precision and recall, calculated as:

$$F1 = 2PR/(P + R). \quad (6)$$

Beside the standard *F1* score we also calculate the *F2* score, which gives twice as much importance to the recall than to the precision:

$$F2 = 5PR=(4P + R). \quad (7)$$

5.4 SBKE Method Results

The results in terms of recall and precision for all 60 documents are in Fig. 2 and in terms of *F1* and *F2* scores are in Fig. 3. Obtained recall and precision exhibit different behavior: recall increases with expanding to SET2 and SET3 while precision decreases. The same is valid for F1 and F2 scores: longer keyword sequences increase F2 score, while the highest F1 score is already achieved on SET2.

Keyword extraction results for 60 individual documents (networks) in terms of average recall (*R*), average precision (*P*), average *F1* and *F2* scores are in the left part of Tab. 1. while recall, precision, *F1* and *F2* scores for the whole HINA document collection are shown in the right part (one integral network). The results of one integral network show

that the expansion of keywords to longer sequences (SET2 and SET3) in terms of precision decreases and that it reaches the highest value for SET1. In contrast, the expansion to SET3 increases the recall. In terms of *F1* score the best result is achieved for the SET2 and *F2* score for the SET3. Hence, we have confirmed the same behavior for the whole collection (one integral network) as in individual document networks (separate networks). Still, it is evident that SBKE yields better results in a larger network (see Tab. 1).

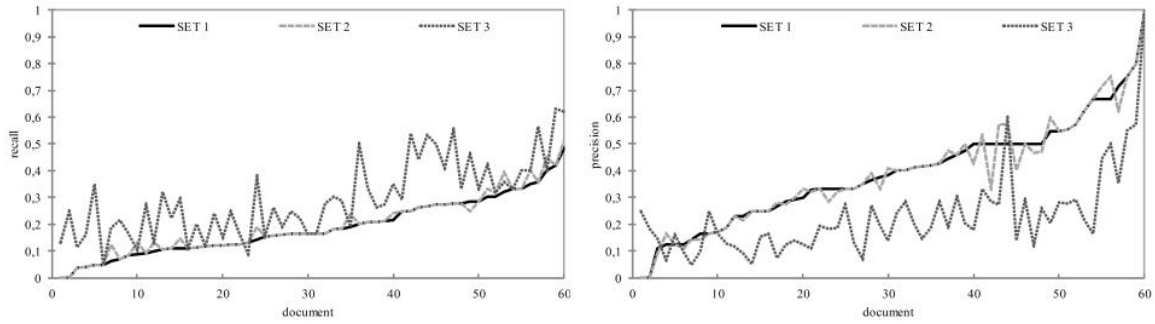


Figure 2: recall and precision for SET1, SET2 and SET3

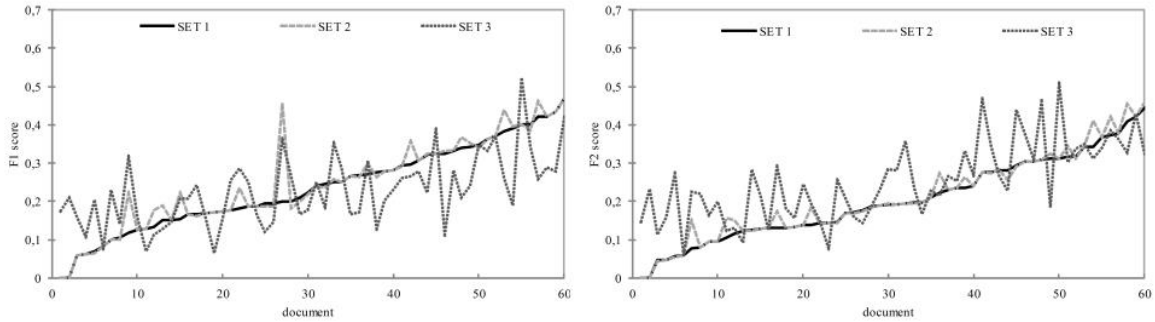


Figure 3: F1 and F2 score for SET1, SET2 and SET3

Table 1: comparison of results for 60 individual (left) and one integral network (right) from all HINA news articles

set	60 individual networks				1 integral network			
	R*[%]	**P***[%]	**F1***[%]	**F2***[%]	**R**[%]	**P**[%]	**F1**[%]	**F2**[%]
SET 1	18.90	**39.35**	23.60	20.48	30.71	**35.80**	33.06	31.61
SET 2	19.74	39.15	**24.76**	21.32	33.46	33.97	**33.71**	33.56
SET 3	**29.55**	22.96	22.71	**25.32**	**60.47**	19.89	28.89	**42.07**

* average value

5.5 Discussion

The reported supervised and unsupervised methods on keyphrase extraction from HINA dataset incorporate linguistic knowledge (POS, MSD) of Croatian language. Mijić at al. [9] reduced the evaluation set to keywords to only top 3 annotators, which had the highest inter-annotator agreement among all 8 annotators and are not directly comparable with ours. Their best results were achieved for n-gram extraction with MSD filtering: F2 was 52.2% and recall was 64.6%. Ahel at al. [3] for one word long

keywords reported a precision of 22% and recall of 3.4%. and F1 value of 15.1% for the extraction of 10 keyphrases. With additional POS filtering they increased F1 to 17.2%, while we obtained F1 score of 24.76% on average on SET2. Bekavac et al. in GPKEX keyphrase extraction method based on genetic programming used the generalized average precision (GAP) measure to evaluate a ranked list of extracted keyphrases. Because of a different evaluation methodology their results are not directly comparable with ours. They used precision at rank (P@10) and recall at rank (R@10). The best results were achieved on strong agreement level R@10 is 30.2%, and P@10, which is 8.2%, while we have obtained the recall of 29.55% and precision of 22.96% on SET3 on average.

We designed our method purely from statistical and structural information encompassed in the source text which is reflected in the structure of the network. Other graph-based approaches presented in [12, 13, 19] show similar results but they incorporate linguistic knowledge in a form of different syntactic filters (POS tagging, stop-words filtering, noun-phrase parsing, etc.) and therefore, are generally more demanding to implement. We have also shown the significant advantage of SBKE method on a larger network (document collection) than on smaller networks (isolated documents). Keyword extraction of one word long candidates – SET1 from a larger network achieved recall slightly higher than 30%, which is in the individual reached yet at SET3. The highest average recall of 32% was obtained for individual documents, while the maximum value in the integral network was almost doubled and was slightly higher than 60%. In terms of F1 and F2 measures, the results of the extracted keyword from a document collection far precede those from individual documents.

The obtained results indicate that our selectivity-based method increases results which are comparable with existing supervised and unsupervised methods, as well as with human annotators. We found that SBKE performance is better if the sequence of keyword candidate is longer (extended from SET2 or SET3). Finally, we concluded that SKBE is more applicable to longer than to shorter texts. Additionally, our method beside keywords, captures personal names and entities as well, which were not marked as keywords and have lowered precision. Capturing names and entities can be of high relevance for tasks like name-entity recognition, text summarization, etc.

6 Conclusion

In this paper, we propose Selectivity-Based Keyword Extraction (SBKE) method as a new unsupervised method for network-based keyword extraction. The node selectivity value is calculated from a directed weighted network as the average weight distribution on the links of the single node and is used in the procedure of keyword candidate ranking and extraction. We proposed keyword expansion to two or three words long keyword sequences determined by the in- or out- weight of the previous or following node. The evaluation was performed on 60 documents from HINA test dataset individually, but also on a large network constructed for the whole document collection.

The experimental results point out that the selectivity-based keyword extraction has a great potential for extracting keywords from larger networks (e.g. collection of documents). We also show that the expansion to three word long keywords achieves the best results in terms of F1 and F2 scores.

Furthermore, the extraction of keywords from one document is comparable with the existing supervised and unsupervised keyphrase extraction methods for Croatian language. The results are comparable especially if we take into account the fact that our approach incorporates no linguistic knowledge, but is derived from pure statistics and the structure of the text is obtained from the network. Since there are no manual annotations required and preprocessing is minimized, fast computing is also an advantage of our selectivity-based method.

In the future work we plan to investigate the SBKE method on: (1) different text types – considering the texts of different length, genre and topics, (2) other languages – tests on standard English and other datasets, (3) new evaluation strategies – considering all flective word forms; considering different matching strategies – exact, fuzzy, part of match, (4) entity extraction – test on whether entities can be extracted from complex networks, (5) text summarization – using SBKE in extraction step in order to identify the most salient elements in text.

7 References

[1] Wu, J-L.; Agogino, A. M. Automating Keyphrase Extraction with Multi-Objective Genetic Algorithms, In Proc. of the 37th Annual Hawaii International Conference on System Science, pages 104-111, HICSS, 2003.

[2] Zhang, Y.;Milios E.; Zincir-Heywood, N. A Comparison of Keyword- and Keyterm-based Methods for Automatic Web Site Summarization, In Technical Report: Papers for the on Adaptive Text Extraction and Mining, pages 15-20, San Jose, 2014.

[3] Ahel, R.;Dalbelo Bašić, B.; Šnajder, J. Automatic Keyphrase Extraction from Croatian Newspaper Articles, In Proc. The Future of Information Sciences, Digital Resources and Knowledge Sharing, pages 207-218, 2009.

[4] Witten, I. H.;Paynter, G. W.; Frank, E.; Gutwin, C.; Nevill-Manning, C.G. Kea: Practical Automatic Keyphrase Extraction, In Proc. of the 4th ACM conference of the Digital Libraries, Berkeley, CA, USA, 1999.

[5] Turney, P. D. Learning to Extract Keyphrases from Text, Technical Report, National Research Council of Canada, Institute for Information Technology, 1999.

[6] Tsatsaronis, G.; Varlamis, I.; Nørvåg, K. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs, ACL 23rd International Conference on Computational Linguistic, pages 1074-1082, Beijing, China, 2010.

[7] Liu, H.; Hu, F. What role does syntax play in a language network?, Europhysic Letters, 83(1):18002, 2008.

[8] Beliga, S.; Meštrović, A.; Martinčić-Ipšić, S. Toward Selectivity Based Keyword Extraction for Croatian News, Submitted on Workshop on Surfacing the Deep and the Social Web, Co-organized by ICT COST Action KEYSTONE (IC1302), Riva del Garda, Trento, Italy, 2014.

[9] Mijić, J.; Dalbelo Bašić, B.; Šnajder, J. Robust Keyphrase Extraction for a Largescale Croatian News Production System, In Proc. of EMNLP, 59-99, 2010.

[10] Bekavac, M.; Šnajder, J. GPKEX: Genetically Programmed Keyphrase Extraction from Croatian Texts, In proceedings of 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 43-47, Sofia, 2013.

[11] Lahiri, S.; Choudhury, S. R. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks, arXiv preprint arXiv:1401.6571, 2014.

[12] Boudin, F. A Comparison of Centrality measures for Graph-Based Keyphrase Extraction, International Joint Conference on Natural Language Processing (IJCNLP), pages 834-838, Nagoya, Japan, 2013.

[13] Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Texts, Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004.
[14] Litvak, M.; Last, M. Graph-Based Keyword Extraction for Single-Document Summarization, Proc. of the workshop on Multi-source Multilingual Information Extraction and Summarization, pages 17-24, Stroudsburg, PA, USA, 2008.
[15] Grineva, M.; Grinev, M.; Lizorkin, D. Extracting Key Terms From Noisy and Multi-theme Documents, In Proc. of the 18th international conference on World wide web, pages 661-670, NY, USA, 2009.
[16] Erkan, G.; Radev, D. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Jour. of Artificial Intelligence Research, 22, 457-479, 2004.
[17] Tsatsaronis, G.; Varlamis, I.; Nørvåg, K. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs, In Proc. of the 23rd International Conference on Computational Linguistic, pages 1074-1082, Stroudsburg, PA, USA, 2010.
[18] Saratlija, J.; Šnajder, J.; Dalbelo Bašić, B. Unsupervised Topic-Oriented Keyphrase Extraction and its Application to Croatian, In Proc. of International Conference on Text, Speech and Dialogue, LNAI 6836, pages 340–347, 2011.
[19] Xie, Z. Centrality Measures in Text Mining: Prediction of Noun Phrases that Appear in Abstracts, In Proc. of 43rd Annual Meeting of the Association for Computational Linguistics, ACL, University of Michigan, USA, 2005.
[20] Wan, X.; Xiao, J. Single Document Keyphrase Extraction Using Neighborhood Knowledge, In Proc.of the 23rd AAAI Conference on Artificial Intelligence, pages 855-860, 2008.
[21] Hagberg, A.; Swart, P.; Chult, D. Exploring Network Structure, Dynamics, and Function Using Networkx, Tech. Report, Los Alamos National Laboratory, 2008.
[22] Masucci, A.; Rodgers, G. Differences between Normal and Shuffled Texts: Structural Properties of Weighted Networks, Advances in Complex Systems, 12(01):113-129, 2009.
[23] Masucci, A.; Rodgers, G. Network Properties of Written Human Language, Physical Rewiew E, 74(2):026102, 2006.
[24] Newman, M. E. J. Networks: An Introduction, Oxford University Press, 2010.
[25] Mihalcea, R. Graph-based Ranking Algorithms for Semantic Extraction, Applied to Text Summarization, In Proc. of 42nd Annual Meeting of the Association for Computational Linguistics, volume (ACL 2004), Barcelona, Spain, 2004.
[26] Frank, E.; Paynter, G. W. Domain-Specific Keyphrase Extraction, In Proc. of the 16th International Joint Conference on Artificial Intelligence, 1999.
[27] Zhang, K.; Xu, H.; Tang, J.; Li, J. Keyword Extraction Using Support Vector Machine, In Proc. of the 7th International Conference on Web-Age, Information Management (WAIM 2006), 2006.
[28] Margan, D.; Martinčić-Ipšić, S.; Meštrović, A. Preliminary report on the structure of Croatian linguistic co-occurrence networks, In Proc. of 5th International Conference on Information Technologies and Information Society, pages 89-96, Slovenia, 2013.

Toward Network-based Keyword Extraction from Multitopic Web Documents

Sabina Šišović, Sanda Martinčić-Ipšić, Ana Meštrović
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{ssisovic, smarti, amestrovic}@uniri.hr

Abstract. *In this paper we analyse the selectivity measure calculated from the complex network in the task of the automatic keyword extraction. Texts, collected from different web sources (portals, forums), are represented as directed and weighted co-occurrence complex networks of words. Words are nodes and links are established between two nodes if they are directly co-occurring within a sentence. We test different centrality measures for ranking nodes - keyword candidates. The promising results are achieved using the selectivity measure. Then we propose an approach which enables extracting word pairs according to the values of the in/out-selectivity and weight measures combined with filtering.*

Keywords. keyword extraction, complex networks, co-occurrence language networks, Croatian texts, selectivity

1 Introduction

Keyword extraction is an important task in the domain of the Semantic Web development. It is a problem of automatic identification of the important terms or phrases in text documents. It has numerous applications: information retrieval, automatic indexing, text summarization, semantic description and classification, etc. In the case of web documents it is a very demanding task: it requires extraction of keywords from web pages that are typically noisy, overburden with information irrelevant to the main topic (navigational information, comments, future announcements, etc.) and they usually contain several topics [3]. Therefore, in keyword extraction from web pages we are dealing with noisy and multitopic datasets.

Various approaches have been proposed for keywords and keyphrases identification (extraction) task. There are two main classes of approaches: supervised and unsupervised. Supervised approaches are based on using machine learning techniques on the manually annotated data [19, 20]. Therefore supervised approaches are time consuming and expensive. Unsupervised approaches may include clustering [7], language modelling [18] and graph-based approaches. Unsupervised approaches may also require different sets of

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

external data, however these approaches are not depended on manual annotation. These approaches are more robust, but usually less precise [2].

A class of graph-based keyword extraction algorithms overcome some of these problems. In graph-based or network-based approaches the text is represented as a network in a way that words are represented as nodes and links are established between two nodes if they are co-occurring within the sentence. The main idea is to use different centrality measures for ranking nodes in the network. Nodes with the highest rank represent words that are candidates for keywords and keyphrases. In [5] an exhaustive overview of network centrality measures usage in the keyword identification task is given.

One of the probably most influential graph-based approaches is the TextRank ranking model introduced by Mihalcea and Tarau in [14]. TextRank is a modification of PageRank algorithm and the basic idea of this ranking technique is to determine the importance of a node according to the importance of its neighbours, using global information recursively drawn from the entire network. However, some recent researches have shown that even simpler centrality measures can give satisfactory results. Boudin [2] and Lahiri et al. [5] compare different centrality measures for keyword extraction task. Litvak and Last [6] compare supervised and unsupervised approach for keywords identification in the task of extractive summarization.

We have already experimented with graph-based approaches for Croatian texts representation. In [12, 13] we described graph-based word extraction and representation from the Croatian dictionary. We used lattice to represent different semantic relations (partial semantic overlapping, more specific, etc.) between words from the dictionary. In [8, 10, 17] we described and analysed network-based representation of Croatian texts. In [10] our results showed that in-selectivity and out-selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. It seems that selectivity measure is able to capture typical word phrases and collocations which are lost during the shuffling procedure. The same holds for English where Masucci and Rodgers [11] found that selectivity somehow captures the specialized local structures in nodes' neighborhood and forms of the morphological structures in text. According to these results, we expected that node selectivity may be potentially important for the text categories differentiation and include it in the set of standard network measures. In [17] we show that the node selectivity measure can capture structural differences between two genres of text.

This was the motivation for further exploration of selectivity for keyword extraction task from Croatian multitopic web documents. We have already analysed the selectivity-based keyword extraction in Croatian news [1]. In this paper we propose an in/out-selectivity based approach combined with filtering to extract keyword candidates from the co-occurrence complex network of text. We design selectivity-based approach as unsupervised, data and domain independent. In its basic form, only the stopwords list is a prerequisite for applying stopwords-filter. As designed, it is a very simple and robust approach appropriate for extraction from large multitopic and noisy datasets.

In Section 2 we present measures for the network structure analysis. In Section 3 we describe datasets and the construction of co-occurrence networks from used text collection. In Section 4 are the results of keyword extraction, and in the final Section 5, we elaborate the obtained results and make conclusions regarding future work.

2 The network measures

This section describes basic network measures that are necessary for understanding our approach. More details about these measures can be found in [11, 15, 16]. In the network, N is the number of nodes and K is the number of links. In weighted language networks every link connecting two nodes i and j has an associated weight w_{ij} that is a positive integer number.

The node degree k_i is defined as the number of links incident upon a node. The in degree and out degree $k_i^{in/out}$ of node i is defined as the number of its in and out neighbours.

Degree centrality of the node i is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$\mathrm{dc}_i = \frac{k_i}{N-1}. \tag{1}$$

Analogously, the in-degree centralities are defined as in-degree of a node:

$$\mathrm{dc}_i^{in} = \frac{k_i^{in}}{N-1}. \tag{2}$$

The out-degree centrality of a node is defined in a similar way. Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest paths between a node and all the other nodes. Let d_{ij} be the shortest path between nodes i and j. The normalised closeness centrality of a node i is given by:

$$\mathrm{cc}_i = \frac{N-1}{\sum_{i \neq j} d_{ij}}. \tag{3}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that traverse through the node i. The normalised betweenness centrality of a node i is given by:

$$\mathrm{bc}_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{4}$$

The strength of a node i is a sum of weights of all links incident with the node i:

$$\mathrm{s}_i = \sum_j w_{ij}. \tag{5}$$

All given measures are defined for directed networks, but language networks are weighted, therefore, the weights should be considered. In the directed network, the in-strength s_i^{in} of the node i is defined as the number of its incoming links, that is:

$$\mathrm{s}_i^{in} = \sum_j w_{ji}. \tag{6}$$

The out-strength is defined in a similar way. The selectivity measure is introduced in [11]. It is actually an average strength of a node. For a node i the selectivity is calculated as a fraction of the node weight and node degree:

$$\mathrm{e}_i = \frac{\mathrm{s}_i}{k_i}. \tag{7}$$

In the directed network, the in-selectivity of the node i is defined as:

$$\mathrm{e}_i^{in} = \frac{\mathrm{s}_i^{in}}{k_i^{in}}. \tag{8}$$

The out-selectivity is defined in a similar way.

3 Methodology

3.1 The construction of co-occurrence networks

Dataset contains 4 collections of web documents written in Croatian language collected from different web sources (portals and forums on different daily topics). The 4 different web sources: business portal Gospodarski list (GL), legislative portal Narodne novine (NN), news portal with forum Index.hr (IN), daily newspaper portal Slobodna Dalmacija (SD).

The first step in networks construction was text preprocessing: "cleaning" special symbols, normalising Croatian diacritics (č, ć, ž, š, dž), and removing punctuations which does not mark the end of a sentence. Commonly, for Croatian which is highly flective Slavic language the lemmatisation and part-of-speech tagging should be performed, but we model our experiment without any explicit language knowledge.

For each dataset we constructed weighted and directed co-occurrence network. Nodes are words that are linked if they are direct neighbours in a sentence. The next step was introducing the networks as weighted edgelists, which contain all the pairs of connected words and their weights (the number of connections between two same words). In the Table 1 there are number of words, number of nodes and number of links per each dataset. We used Python and the NetworkX software package developed for the construction, manipulation, and study of the structure, dynamics, and functions of complex networks [4].

3.2 The selectivity-based approach

The goal of this experiment is to analyse the selectivity measure in the automatic keyword extraction task. First, we compute centrality measures for each node in all 4 networks: in-degree centrality, out-degree centrality, closeness centrality, betweenness centrality and selectivity centrality. Then we rank all nodes (words) according to the values of each of these measures, obtaining top 10 keyword candidates automatically from the network.

In the second part of our experiment we compute in-selectivity and out-selectivity for each node in all 4 networks. The nodes are then ranked according to the highest in/out-selectivity values. Then, for every node we detect neighbour nodes with the highest weight. For the in-selectivity we isolate one neighbour node with the highest outgoing link weight. For the out-selectivity we isolate one neighbour node with the highest ingoing link weight. The result of in/out-selectivity extraction is a set of ranked word tuples.

The third part of our approach consider applying different filters on the in/out-selectivity based word tuples. The first is the stopwords-filter: we filter out all tuples that contain stopwords. Stopwords are a list of the most common, short function words which

Dataset	GL	NN	IN	SD
Number of words	199 417	146 731	118 548	44 367
Number of nodes N	27727	13036	15065	9553
Number of links K	105171	55661	28972	25155

Table 1: The number of words, number of nodes and number of links for all 4 datasets

	selectivity	in-degree	out-degree	closeness	betweenness
1.	mladićevi (joungsters)	i (and)	i (and)	je (is)	i (in)
2.	pomlatili (beaten)	u (in)	je (is)	i (and)	je (is)
3.	seksualnog (sexual)	je (is)	u (in)	se (self)	u (in)
4.	policijom (police)	na (on)	na (on)	da (that)	na (on)
5.	uhićeno (arrested)	da (that)	se (self)	su (are)	se (self)
6.	skandala (scandal)	za (for)	za (for)	to (it)	za (for)
7.	podnio (submitted)	se (self)	su (are)	a (but)	da (that)
8.	obožavatelji (fans)	a (but)	da (that)	će (will)	su (are)
9.	sata (hour)	su (are)	s (with)	samo (only)	a (but)
10.	Baskiji (Baskia)	s (with)	od (from)	ali (but)	s (with)

Table 2: Top ten words from the dataset IN ranked according to the selectivity, in/out-degree, closeness and betwenness

do not carry strong semantic properties, but are needed for the syntax of language (pronouns, prepositions, conjunctions, abbreviations, interjections,...). The second is the high-weights-filter: from the in/out-selectivity based word tuples we chose only those tuples that have the same values for the selectivity and weight. The third filter is the combination of the first two filters.

4 Results

Initially, we analyse 4 networks constructed for each dataset. The top 10 ranked nodes with the highest values of the selectivity, in degree, out degree, closeness and betwenness measures for datasets IN, GL, SD and NN are shown in the Tables 2,3,4 and 5. It is obvious that top 10 ranked words according to the in/out degree centrality, closeness centrality and betwenness centrality are stopwords. It can be also noticed that centrality measures return almost identical top 10 stopwords. However, the selectivity measure ranked only open-class words: nouns, verbs and adjectives. We expect that among these highly ranked words are keyword candidates.

Furthermore, we analyse selectivity measure in details. Since texts are better represented as directed networks [9], we analyse words with in-selectivity and out-selectivity measure separately. We extract word-tuple: the word before for in-selectivity and the word after for out-selectivity that has the highest value of the weight. In Table 6 are shown ten highly ranked in/out-selectivity based word-tuples together with the values of in/out-selectivity and weight.

Hence, we extract the most frequent word-tuples which are possible collocations or phrases from the text. We expect that among these highly ranked word-tuples are keyword

	selectivity	in degree	out degree	closeness	betweenness
1.	stupastih (cage)	i (and)	i (and)	i (and)	i (and)
2.	populaciju (population)	u (in)	u (in)	se (self)	u (in)
3.	izdanje (issue)	na (on)	je (is)	je (is)	je (is)
4.	online (online)	je (is)	se (self)	su (are)	na (on)
5.	webshop (webshop)	ili (or)	na (on)	a (but)	se (self)
6.	matrica (matrix)	a (but)	ili (or)	ili (or)	ili (or)
7.	pretplata (subscription)	se (self)	su (are)	to (it)	a (but)
8.	časopis (journal)	za (for)	za (for)	bolesti (disease)	za (for)
9.	oglasi (ads)	od (from)	od (from)	da (that)	su (are)
10.	marketing (marketing)	su (are)	a (but)	biljke (plants)	od (from)

Table 3: Top ten words from the dataset GL ranked according to the selectivity, in/out-degree, closeness and betwenness

	selectivity	in-degree	out-degree	closeness	betweenness
1.	seronjo (bullshitter)	i (and)	i (and)	i (and)	i (and)
2.	Splitu (Split)	u (in)	je (is)	je (is)	je (is)
3.	upišite (fill-in)	je (is)	u (in)	svibanj (May)	u (in)
4.	uredniku (editor)	komentar (comment)	se (self)	se (self)	se (self)
5.	ekrana (monitor)	na (on)	svibanj	ali (but)	na (on)
6.	crkvu (church)	se (self)	na (on)	a (but)	od (from)
7.	supetarski (Supetar)	za (for)	za (for)	će (will)	za (for)
8.	vijesti (news)	a (but)	da (that)	to (it)	a (but)
9.	zaradom (earning)	svibanj (May)	ne (ne)	još (more)	svibanj
10.	Jović (Jović)	od (from)	a (but)	pa (so)	to (it)

Table 4: Top ten words from the dataset SD ranked according to the selectivity, in/out-degree, closeness and betwenness

	selectivity	in-degree	out-degree	closeness	betweenness
1.	novine (newspaper)	i (and)	i (and)	i (and)	i (and)
2.	temelju (based on)	u (in)	u (in)	ili (or)	u (in)
3.	manjinu (minority)	za (for)	je (is)	je (is)	za (for)
4.	srpsku (Serbian)	na (on)	za (for)	se (self)	ili (or)
5.	sladu (harmony)	ili (or)	se (self)	da (that)	na (on)
6.	snagu (strength)	iz (from)	ili (or)	usluga (service)	je (is)
7.	osiguranju (insurance)	te (and)	na (on)	zakona (law)	se (self)
8.	narodnim (national)	je (is)	o (on)	a (but)	o (on)
9.	novinama (newspaper)	se (self)	te (and)	skrbi (welfare)	te (and)
10.	kriza (crisis)	s (with)	članak (article)	HRT-a (HRT-a)	iz (form)

Table 5: Top ten words from the dataset NN ranked according to the selectivity, in/out-degree, closeness and betwenness

	in-selectivity			out-selectivity		
	word tuple	e^{in}	w	word tuple	e^{out}	w
1.	narodne **novine**	326	326	**srpsku** nacionalnu	222	222
2.	na **temelju**	317	317	**nacionalnu** pripadnost	183	1
3.	nacionalnu **manjinu**	275	2	**ovjesne** jedrilice	159	159
4.	za **srpsku**	222	222	**narodnim** novinama	129	129
5.	u **skladu**	202	202	**narodne** jazz	111	1
6.	na **snagu**	172	172	**manjinu** gradu	78	1
7.	o **osiguranju**	134	43	**ovoga** sporazuma	72	1
8.	u **narodnim**	129	129	**crvenog** kristala	72	3
9.	narodnim **novinama**	129	129	**skladu** provjeriti	67	1
10.	crvenog **križa**	99	2	**oružani**h sukoba	58	4

Table 6: Top ten highly ranked in/out-selectivity based word-tuples for the NN dataset

	in-selectivity			out-selectivity		
	word tuple	e^{in}	w	word tuple	e^{out}	w
1.	narodne **novine**	326	326	**srpsku** nacionalnu	222	222
2.	nacionalnu **manjinu**	275	2	**nacionalnu** pripadnost	183	1
3.	narodnim **novinama**	129	129	**ovjesne** jedrilice	183	1
4.	crvenoga **križa**	99	2	**narodnim** novinama	129	129
5.	jedinicama **regionalne**	65	1	**narodne** jazz	111	1
6.	nacionalne **manjine**	61	61	**manjinu** gradu	78	1
7.	rizika **snaga**	57	1	**ovoga** sporazuma	72	1
8.	medije **ubroj**	47	1	**crvenog** kristala	72	3
9.	crveni **križ**	42	42	**skladu** provjeriti	67	1
10.	uopravni **spor**	41	41	**oružanih** sukoba	58	4

Table 7: Top ten highly ranked in/out-selectivity based word-tuples without stopwords for the NN dataset

in-selectivity		out-selectivity	
word tuple	e^{in}=w	word tuple	e^{out}=w
na **temelju** (based on)	317	**ovjesne** jedrilice (hangh glider)	159
za **srpsku** (for Serbian)	222	**narodnim** novinama (Nat. news.)	129
u **skladu** (according to)	202	**sjedištem** u (headquarter in)	55
na **snagu** (into effect)	172	**objavit** će (will be bublished)	53
u **narodnim** (in national)	129	**republici** Hrvatskoj (Croatia)	52
narodnim **novinama** (Nat. news.)	129	**albansku** nacionalnu (Alb. nat.)	52
i **dopunama** (and amendments)	68	**republika** Hrvatska (Croatia)	49
nacionalne **manjine** (nat. minority)	61	**oplemenjivačkog** prava (noble law)	45
sa **sjedištem** (with headquarter)	55	**madjarsku** nacionalnu (Hung. nat.)	40

Table 8: Top ten highly ranked in/out-selectivity based word-tuples with equal in/out-selectivity and weight for the NN dataset

in-selectivity word tuple	out-selectivity word tuple
narodne **novine** (National newspaper)	**srpsku** nacionalnu (Serbian national)
narodnim **novinama** (Nat. newspapers)	**ovjesne** jedrilice (hangh glider)
nacionalne **manjine** (nat. minority)	**narodnim** novinama (Nat. newspapers)
crveni **križ** (red cross)	**republici** hrvatskoj (Republic of Croatia)
upravni **spor** (administrative dispute)	**albansku** nacionalnu (Albanian national)
ovjesnom **jedrilicom** (hangh glider)	**republika** hrvatska (Republic of Croatia)
elektroničke **medije** (electronic media)	**oplemenjivačkog** prava (noble law)
nacionalnih **manjina** (national minority)	**madjarsku** nacionalnu (Hungarian nat.)
domovinskog **rata** (Homeland War)	**romsku** nacionalnu (Romany national)
Ivan **Vrljić** (Ivan Vrljić)	**nadzorni** odbor (supervisory board)

Table 9: Top ten highly ranked in/out-selectivity based word-tuples with equal in/out-selectivity and weight without stopwords for the NN dataset

candidates. Due to limited space, we show results only for the NN dataset, but other datasets raised similar results.

In Table 6 there are word-tuples which contain stopwords, especially for the in-selectivity based ranking.Therefore we use stopwords-filter defined in the previous section as shown in Table 7. Now we obtain more open class keyword candidates from highly ranked word-tuples.

In Table 8. there are 10 highly ranked word-tuples for the NN dataset with the high-weights-filter applied. Using this approach some new keyword candidates appear in the ranking results.

In Table 9. there are 10 highly ranked word-tuples from the NN dataset with the both filters applied. According to our knowledge about the content of the dataset, these two filters derived the best results.

5 Conclusion and discussion

We analyse network-based keyword extraction from multitopic Croatian web documents using selectivity measure. We compare keyword candidate words rankings with selectivity and three network-based centrality measures (degree, closeness and betwenness). The selectivity measure gives better results because centrality-based rankings select only stopwords as the top 10 ranked words. Furthermore, we propose extracting the highly connected word-tuples with the highest in/out-selectivity values as the keyword candidates. Finally, we apply different filters (stopwords-filter, high-weights-filter) in order to keyword candidate list.

The first part of analysis can raise some considerations regarding the selectivity measure. The selectivity measure is important for the language networks especially because it can differentiate between two types of nodes with high strength values (which means words with high frequencies). Nodes with high strength values and high degree values would have low selectivity values. These nodes are usually stopwords (conjunctions, prepositions,...). On the other side, nodes with high strength values and low degree values would have high selectivity values. These nodes are possible collocations, keyphrases and names that appear in the texts. It seems that selectivity is insensitive to stopwords (which are the most frequent words) and therefore can efficiently detect semantically rich open

class words from the network.

Furthermore, since we modelled multitopic datasets the keyword extraction task is even more demanding. From the results of this preliminary research it seems that the selectivity has a potential to extract keyword candidates without preprocessing (lemmatisation, POS tagging) from multitopic sources.

There are several drawbacks in this reported work: we did not perform the classical evaluation procedure because we did not have annotated data and we conducted analysis only on Croatian texts.

For the future work we plan to evaluate our results on different datasets in different languages. Furthermore, it seems promising to define an approach that can extract a sequence of three or four neighbouring words based on filtered word-tuples. We also plan to experiment with lemmatised texts. Finally, in the future we will examine the effect of noise to the results obtained from multitopic sources.

References

[1] S. Beliga, A. Meštrović and S. Martinčić-Ipšić. Toward Selectivity Based Keyword Extraction for Croatian News. Submitted on Workshop on Surfacing the Deep and the Social Web, Co-organised by ICT-COST Action KEYSTONE (IC1302), Riva Del Garda, Trento, Italy, 2014.

[2] F. Boudin. A comparison of centrality measures for graph-based keyphrase extraction. International Joint Conference on Natural Language Processing. pp. 834–838, (2013)

[3] M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. ACM 18th conference on World wide web, pp.661–670, (2009)

[4] A. Hagberg, P. Swart, and D. Chult. Exploring network structure, dynamics, and function using networkx. (2008)

[5] S. Lahiri, S.R. Choudhury, and C. Caragea. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks. arXiv:1401.6571, (2014)

[6] M. Litvak and M. Last. Graph-based keyword extraction for single-document summarization. ACM Workshop on Multi-source Multilingual Information Extraction and Summarization. pp.17–24, (2008)

[7] Z. Liu, P. Li, Y. Zheng, and M. Sun. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Volume 1-Volume 1, pp. 257-266 (2009)

[8] D. Margan, S. Martinčić-Ipšić and A. Meštrović. Network Differences Between Normal and Shuffled Texts: Case of Croatian. Studies in Computational Intelligence, Complex Networks V. Vol.549. Italy, pp. 275–283 (2014)

[9] D. Margan, S. Martinčić-Ipšić, and A. Meštrović. Preliminary report on the structure of Croatian linguistic co-occurrence networks. 5th International Conference on Information Technologies and Information Society (ITIS), Slovenia, 89–96 (2013)

[10] D. Margan, A. Meštrović and S. Martinčić-Ipšić. Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts. IEEE MIPRO 2014, Croatia, pp.1819–1823 (2014)

[11] A. Masucci and G. Rodgers. Differences between normal and shuffled texts: structural properties of weighted networks. Advances in Complex Systems, 12(01):113–129 (2009)

[12] A. Meštrović and M. Čubrilo. Monolingual dictionary semantic capturing using concept lattice. International Review on Computers and Software 6(2):173–184 (2011)

[13] A. Meštrović. Semantic Matching Using Concept Lattice. In Proc. of Concept Discovery in Unstructured Data, Katholieke Universiteit Leuven, pp. 49–58 (2012)

[14] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. ACL Empirical Methods in Natural Language Processing, (2004)

[15] M. E. J. Newman. Networks: An Introduction. Oxford University Press.(2010)

[16] T. Opsahl, F. Agneessens and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3): 245–251 (2010)

[17] S. Šišović, S. Martinčić-Ipšić and A. Meštrović. Comparison of the language networks from literature and blogs. IEEE MIPRO 2014, Croatia, pp.1824–1829, (2014).

[18] T. Tomokiyo, and M. Hurst. A language model approach to keyphrase extraction. In Proc. of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment,Volume 18, pp. 33–40 (2003)

[19] P. D. Turney. Learning algorithms for keyphrase extraction. 2(4):303–336 (2000)

[20] I. H. Witten et al. Nevill-Manning. Kea: practical automatic keyphrase extraction. In Proc. of the fourth ACM conference on Digital libraries, pp. 254-255 (1999)

Toward a Complex Networks Approach on Text Type Classification

Domagoj Margan, Ana Meštrović, Marina Ivašić-Kos, Sanda Martinčić-Ipšić
Department of Informatics University of Rijeka Radmile Matejčić 2, 51000 Rijeka, Croatia
{dmargan, amestrovic, marinai, smarti}@uniri.hr

Abstract: *The growing amount of text electronically available has placed text type classification among the most exciting issues in the field of exploratory data mining. This talk presents an preliminary approach to text type classification by features of linguistic co-occurrence networks. Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. The aim of our work-in-progress presented in this talk is to investigate the idea of replacing the standard natural language processing feature sets with linguistic network measures for the purpose of text type clas- sification. This talk tackles the problem of binary classification of two different text types. Our dataset is consisted of 150 equal-sized Croatian texts divided in two classes: 75 liter- ature texts and 75 blog texts. Literature texts represent segments from 7 different books written in or translated to Croatian language, while blog texts are collected from two very popular Croatian blogs. The trait which prompted us to do the classification of this par- ticular text types is the linguistic distinction between book and blog. We constructed 150 different co-occurrence networks (one for each text in the dataset), all weighted and di- rected. Words are nodes linked if they are co-occurring as neighbors to each other in a sentence. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a text. For each network we computed a set of 10 measures (number of components, average degree, average path length, clustering coeffi- cient, transitivity, degree assortativity, density, reciprocity, average in-selectivity, average out-selectivity), which are used as feature set for classification. All features are rescaled to [0 − 1] in order to make them independent of each other. We preformed a series of clas- sification experiments using various types of classification algorithms and methods (sup- port vector machine, classification trees, Naive Bayes, k-nearest neighbor, LDA, QDA). The performance of each classifier was evaluated with corresponding methods, such as misclassification error measures, confusion matrices and receiver operating characteristic curves. All classification experiments show very good classification accuracy, while the average in- and out- selectivity measures act as the most useful features in predicting the correct text type and reducing the misclassification rate. Precision and recall measures and ROC curves indicate that the node selectivity measures are the only measures from the feature set that can capture the structural differences between two classes of networks.*

Key Words*: complex networks, linguistic co-occurrence networks, text type classification, document classification…*

Machine Learning in Classification of News Portal Articles

Renato Barišić
University College Algebra
Ilica 242, 10000 Zagreb, Croatia
renato.barisic@racunarstvo.hr

Jože Bučar
Faculty of Information Studies
Ulica talcev 3, 8000 Novo mesto, Slovenia
joze.bucar@fis.unm.si

Abstract: *Machine learning algorithms can find out how to perform tasks by generalizing from examples. As more data becomes available, more ambitious problems can be tackled and, as a result of that, machine learning is widely used in computer science and other fields. Text classification is one of the booming areas in research because of the availability and the necessity to process huge amount of electronic data in the form of news articles, research articles, email messages, blogs, web pages, etc. This paper shows algorithms and classification methods for processing news portal articles using machine learning software Weka and Naive Bayes, k-Nearest Neighbour, Support Vector Machines, Decision Rules and Decision Trees (J48/C4.5, CART) classifiers.*

Key words: *algorithm, article, classification, machine learning, news portal, Weka*

1. Introduction

Machine learning addresses the question of how to build computer programs that improve their performance at some task through experience. Machine learning algorithms have proven to be of great practical value in a variety of application domains.

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Many researchers also think it is the best way to make progress towards human-level.

2. Machine learning algorithms and classifiers

In the terminology of machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering or cluster analysis, and involves grouping data into categories based on some measure of inherent similarity (e.g. the distance between instances, considered as vectors in a multi-dimensional vector space).

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term classifier sometimes also refers to the mathematical function, implemented by a classification algorithm, which maps input data to a category. In this paper well known classifiers were used: Naive Bayes, *k*-Nearest Neighbour, Support Vector Machines, Decision Rules, Decision Trees – J48/C4.5, Decision Trees – CART.

2.1 Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines [7].

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests [3].

2.2 *k*-Nearest Neighbour

In pattern recognition, the *k*-Nearest Neighbours algorithm (or *k*-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k*-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its *k* nearest neighbours (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbour.
- In *k*-NN regression, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbours.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms. A shortcoming of the *k*-NN algorithm is its sensitivity to the local structure of the data.

2.3 Support Vector Machines

In machine learning, support vector machines (SVM, also support vector networks) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis.

Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

2.4 Decision Rules

Large decision trees are difficult to understand because each node appears in a specific context established by the outcomes of tests at the antecedent nodes. The work of Rivest [8] presents a new representation, a decision list, which generalizes decision trees. The advantage of this representation is modularity and consequently interpretability: each rule is independent of the others, and can be interpreted in isolation of the others. Rule sets take advantage of not being hierarchically structured, so concept descriptions can be updated or removed when becoming out–of–date without hardly affecting the learning efficiency.

2.5 Decision Trees – J48/C4.5

Classifier J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is an algorithm used to generate a decision tree developed by [8]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub-lists.

2.6 Decision Trees – CART

Decision tree learning is a method commonly used in data mining [9]. Decision trees used in data mining are of two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. [2]. Trees used for regression and trees used for classification have some similarities but also some differences, such as the procedure used to determine where to split.

3. Materials and methods

Research was based on 200 news articles from four Croatian news portals.

Most popular news portals in Croatia are (in alphabetic order):

- 24 sata, http://www.24sata.hr/
- Dnevnik.hr, http://www.dnevnik.hr/
- Jutarnji list, http://www.jutarnji.hr/
- Vecernji list, http://www.vecernji.hr/

From every one of them 50 articles were randomly taken and content of each article (just plain text) is copied in the text file. After 200 articles were collected and saved, every article needed to be carefully read and, according to human sentiment analysis, saved in one of corresponding folders named "positive", "negative" and "neutral". At the end of sentiment analysis 200 articles were distributed in three different folders.

For further data processing Weka 3.6 was used. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License.

Using Weka's SimpleCLI tool and weka.core.converters.TextDirectoryLoader class, articles from abovementioned three folders were converted in one .arff file. The StringToWordVector filter was used and, for the purpose of this paper, list of Croatian stopwords were used organized as expected by the StringToWordVector filter.

For the last step, a few necessary manual clean-up actions needed to be done. After these final actions, 514 attributes were left as baseline for planned classifications.

4. Results and discussion

Table 1 shows results of classifiers applications as a summary data of corresponding classifier output for:

- Naive Bayes (bayes.NaiveBayes)
- *k*-Nearest Neighbour (lazy.IBk)
- Support Vector Machines (functions.SMO)
- Decision Rules (rules.DecisionTable)
- Decision Trees – J48/C4.5 (trees.J48)
- Decision Trees – CART (trees.SimpleCart)

For all classifiers option Cross-validation is set to Folds = 5.

Table 1: Summary data from classifier outputs

Classifier	Correctly Classified		Incorrectly Classified		Kappa statistics	Mean absolute error
	Instances	%	Instances	%		
Naive Bayes (bayes.NaiveBayes)	138	69	62	31	0,515	0,2162
k-Nearest Neighbour (lazy.IBk)	90	45	110	55	0,141	0,3681
Support Vector Machines (functions.SMO)	135	67,5	65	32,5	0,4957	0,3222
Decision Rules (rules.DecisionTable)	102	51	98	49	0,1825	0,3903
Decision Trees – J48/C4.5 (trees.J48)	105	52,5	95	47,5	0,2455	0,3402
Decision Trees – CART (trees.SimpleCart)	110	55	90	45	0,2748	0,3431

According to the summarized data, the best results are obtained by the Naive Bayes classifier and second best are obtained by the Support Vector Machines classifier. These results are expected because Naive Bayes is known as the method of first choice for text data classification and Support Vector Machines classifier is also known as helpful and one of the top algorithms for text classification problems.

Average results obtained by Decision Trees (CART and J48/C4.5) and Decision Rules classifiers are a little bit unexpected and are probably result of relatively small training set and because of just one fifth of positively labelled articles. Worst results are obtained by *k*-Nearest Neighbour classifier.

It is obvious that all results, even the best ones, obtained by Naive Bayes and SVMs, gave less than expected percentage of correctly classified instances. Reason for that, fairly certain, was small number of selected news articles and just 21% positively sorted articles during sentiment analysis. As could be seen in Figure 1, 75 articles (37%) were labelled "negative", 84 articles (42%) were labelled "neutral" and just 41 articles (21%) were labelled "positive".

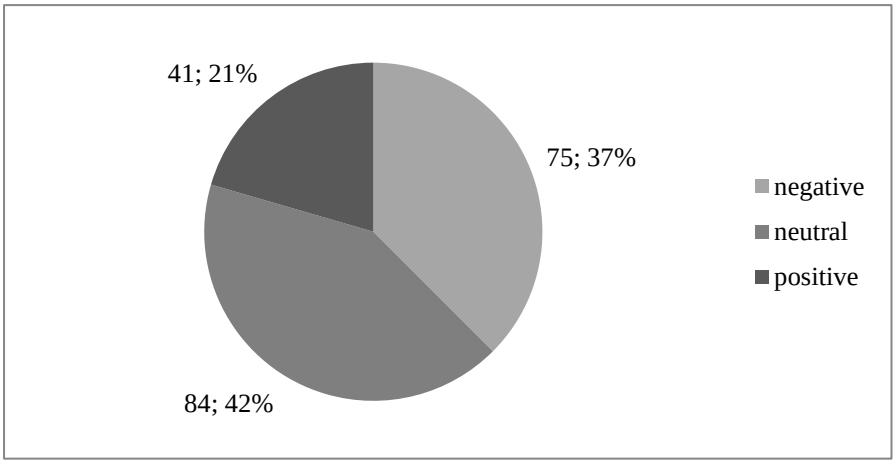


Figure 1: News articles count and sentiment analysis summary

Given that news articles from four news portals are chosen randomly it is obvious that most of contemporary news articles are written with negative or neutral connotation and just 21% of them leaves positive feeling.

5. Conclusion

This paper has shown results of classification using machine learning software Weka 3.6 and well known classifiers: Naive Bayes (bayes.NaiveBayes), *k*-Nearest Neighbour (lazy.IBk), Support Vector Machines (functions.SMO), Decision Rules (rules.DecisionTable), Decision Trees – J48/C4.5 (trees.J48) and Decision Trees – CART (trees.SimpleCart). Classification was conducted on 200 randomly chosen news articles from four Croatian news portals: 24 sata, Dnevnik.hr, Jutarnji list and Vecernji list.

The best classification results are obtained by the Naive Bayes classifier and Support Vector Machines classifier while other classifiers outputs were worst. The results are acceptable but better results are expected for all mentioned classifiers.

Probable reason for worse results was small number of selected news articles and just 21% positively sorted articles during sentiment analysis. For further research it is necessary to use at least one thousand news articles and to strive for more balanced distribution of articles into categories despite the obvious neutral or negativistic aspiration of modern newspapers and news portals.

6. Bibliography

[1] AYODELE, TAIWO OLADIPUPO (2010) Types of Machine Learning Algorithms. *New Advances in Machine Learning*, Yagang Zhang (Ed.), ISBN: 978-953-307-034-6, InTech, Available from: http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms (05.08.2014).
[2] BREIMAN, LEO; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. (1984) *Classification and regression trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN: 978-0-412-04841-8.
[3] CARUANA, RICH and NICULESCU-MIZIL, ALEXANDRU (2006) An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*. Pittsburg: ACM. ISBN: 978-1-59593-383-2.
[4] CLARK, P. and NIBBET, T. (1989) The CN2 induction algorithm. *Machine Learning*, 3, p. 261-283.
[5] MITCHELL, TOM (1997) *Machine Learning.* Boston: McGraw Hill.
[6] QUINLAN, JOHN ROSS (1993) *C4.5: Programs for Machine Learning*. Burlington: Morgan Kaufmann Publishers.
[7] RENNIE, JASON D. M.; SHIH, LAWRENCE; TEEVAN, JAIME; KARGER, DAVID R. (2003) *Tackling the poor assumptions of Naive Bayes classifiers*. Cambridge: MIT Artificial Intelligence Laboratory.
[8] RIVEST, R. (1987) Learning Decision Lists. *Machine Learning*, 2, p. 229-246.
[9] ROKACH, LIOR; MAIMON, O. (2008) *Data mining with decision trees: theory and applications*. Singapore: World Scientific Publishing. ISBN: 978-9812771711.

An approach to predict malicious threats

Andrej Dobrovoljc
Faculty of Information Studies
University of Novo mesto
Sevno 13, 8000 Novo mesto, Slovenia
andrej.dobrovoljc@fis.unm.si

Abstract: *Risk assessment of information systems depends on recognized system vulnerabilities and detected threats. From the security assurance point of view it is essential to discover and remove vulnerabilities as soon as possible. Therefore, various Vulnerability Discovery Models (VDM) were proposed for making predictions about discoveries in software products. Since vulnerabilities without threats are not harmful, we should pay attention to prediction of malicious threats as well. We did not find any threat prediction models in the existing literature. Authors of malicious attacks are always people with bad intentions. Therefore, we should understand their behaviour and detect the moment when the system becomes attractive to them. Our intention in this paper was to uncover possibilities for making such predictions. We describe some possible approaches, which will be deeply examined in further studies.*

Key Words*: threat, discovery, risk, vulnerability, attacker, TAM*

1 Introduction

It is very difficult to answer the question how safe is our information system (IS). Individuals daily discover new and unknown IS vulnerabilities, what allows new attacks and consequently sustain constant threat. With the IS globalization vulnerability discovery became extremely popular. Companies were forced to open their IS to the web in order to keep their competitiveness. Trade via internet and all sorts of financial transactions enormously increased in the last decade as well. All these changes attract criminals, who want to earn something out of it. The first step on the way to the successful exploitation of web applications or protocols is to find their vulnerabilities. By using them, attackers can develop and use various attacks against IS, what leads to huge economical and private financial damage.

Only highly skilled and innovative individuals can discover vulnerabilities. Due to this fact, the knowledge about undisclosed vulnerabilities became a respected merchandise. Today, we are aware of a real vulnerability market where the customers are criminal groups, terrorists, governments, specific private companies and other special groups. In order to avoid such trading some initiatives require better quality of software products regarding their security. Some others demand severe legislation against vendors, which would define their responsibility for security. The vulnerability market is a fact and we have to find some other preventive actions. We should not overlook the other important types of vulnerabilities. Information system does not consists only of software but also of the other types of components. Among them people are the weakest link and vulnerable to various malicious threats such as social engineering attack and the like. Therefore,

some specific risk mitigation mechanisms arose in the security ecosystem, which are mainly focused to vulnerability discovery, elimination, prediction and prevention.

However, security risks do not depend only on vulnerabilities. Until threat appears, there is no risk. Therefore, it is also important to discover possible future threats around the information system. Behind every malicious threat is a human. The question that we address in this paper is how we can anticipate the appearance of future malicious threats. The answer to this question can improve the risk assessment process, since it would allow the IS owner to mitigate risk in advance. In our study, we focused in attackers' behavior as well as on their trails, which are recorded in various databases.

The remainder of this paper is structured as follows. Section two describes existing approaches for proactive security assurance. In the third section the opportunities for malicious threats detection are given. In the final section, a paper summary is given with a short description of main ideas and plans for future work.

2 Related work

Information system (IS) is a combination of hardware, software, infrastructure, data, procedures and trained personnel. With each of the components, we can bind specific risks that exist as a result of existence of their vulnerabilities. We talk about [1]:

- hardware vulnerabilities (natural catastrophes, physical deterioration),
- software vulnerabilities (software misusage such as XSS, Buffer Overflow),
- data vulnerability (obsolete data formats),
- organizational vulnerabilities (mistakes and deficient procedures, user attitudes, information security culture),
- human vulnerabilities (specific behavior, limited perception etc.).

Several approaches exist for making predictions on vulnerability discoveries for individual IS components:

- hardware: MTBF (Mean Time between Failures), MTTF (Mean Time to Failure)
- software: VDM (Vulnerability Discovery Models) [2]
- human: NLP (Neuro-linguistic Programming), TA (transactional analysis) [3]

Anticipating software vulnerability discoveries is not a completely new challenge. However, it became extremely important in the last decade with the development of web applications and technologies. Vulnerability Discovery Model (VDM) shall predict the time and the frequency of future discoveries. This information allows the software producers to acquire needed resources in order to fix defects on time.

Several VDM models have been proposed in last years. According to their prediction approach, we separate them into two categories. *Effort-based models* are based on the number of product users. They are difficult to realize due to the lack of the needed data. Namely, the number of the product users is constantly changing and there are no accurate records about it [4]. The second category are *Time-based models.* Among them is an Alhazmi-Malaiya Logistic model (AML) the most accurate one [5]. It is based on a logistic function and on a quite simple assumption. In an early phase, when the product enters the market, it has a few users. Simultaneously with increased popularity grows also the number of users. At the end of its life cycle, the number of users declines. According to the findings of the authors of AML model, vulnerability discoveries follow the same logistic function. See the Fig.1.

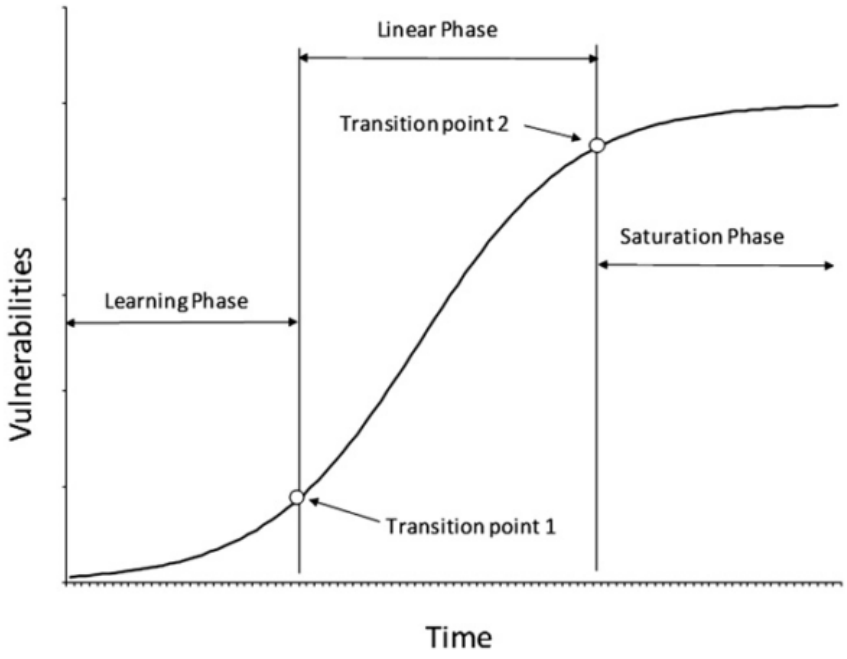


Figure 1 Vulnerability discoveries follow the logistic function

Proposed models only partly consider the factors that affect the discovery process. Besides the number of product users, we have to take into account also the vulnerability management processes of software producers and the learning process of hackers [6].

Hackers (black and white ones) are more interested to discover vulnerabilities on the products, with the highest potential benefit for them. Generally, these are the most popular products on the market, what we can measure with the number of their users [7]. The most dangerous are undiscovered and undisclosed vulnerabilities because software producers are not aware of them. There is simply no protection against attacks based on exploitation of such vulnerabilities

The existence of black and white market proves that the malicious threats exist. When a software producer launches a completely new product, it is unlikely that it would be attractive to attackers at once. Malicious threats do not exist at the very beginning and something has to give them a birth. We believe that some key factors exists, which have an impact on the process of creation of new malicious threats. Unfortunately, in the existing literature there are no models for predicting threats. Therefore, we place the following research questions:

RQ1: Which models exist for predicting people behavior?

RQ2: Which data can be used to estimate the time, when the software product became attractive to the threat agents?

Answers to these research questions will help us develop the model for detecting malicious threats and hopefully to predict them in advance by using planned changes within the IS and its user environment.

3 Opportunities for making predictions about malicious threats

We make predictions on some existing facts or estimates. For predictions of malicious threats we need an appropriate definition of threat. In general, threat can be defined as a function of motivation, the expected impact for the attacker, his or her capability to carry out the attack, and the opportunities for the realization of an attack [8].

The impact is directly related to the asset value, because IS assets are those targets within the IS, which are interesting to attackers and consequently they represent harm to the IS owner. When we combine the definition of threat with the one for security risk, we get the model on the Fig. 2.

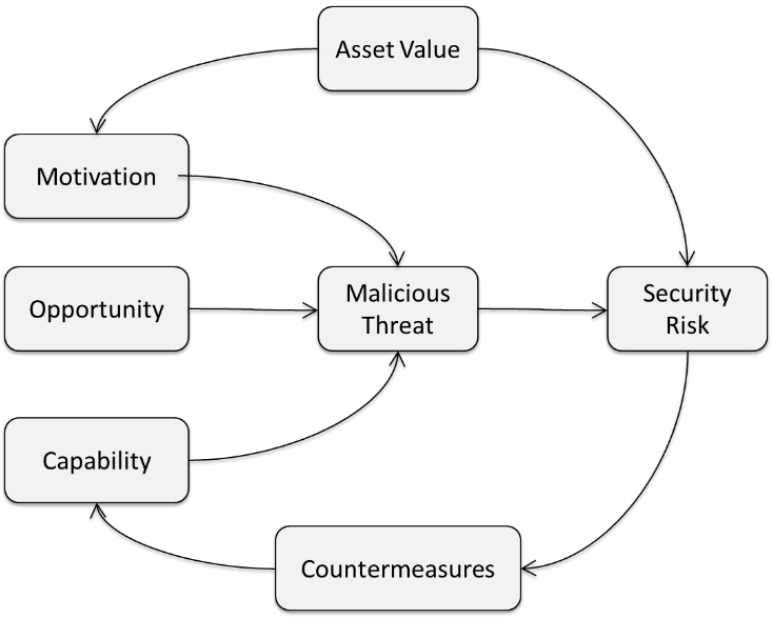


Figure 2 Security risk model for malicious threats

Malicious threat appears when all threat components are available. Behind the malicious threat is always a human. Therefore, we should somehow understand his or her behavior. In case of threat agent, we can speak about the misuse or abuse of information technology. On the field of IT/IS, several models have been developed to assess the acceptance and use of technology. Davis [9] proposed the technology acceptance model (TAM) with the following key concepts:

- *perceived ease of use (PEOU)*: degree to which a person believes that the system is easy to use,
- *perceived usefulness (PU)*: degree to which a person believes that the system will be useful in performing his or her job,
- *behavioral intention (BI)*: degree to which a person has formulated conscious plans to perform or not perform some specified future behavior.

TAM is a flexible model. It has been extended several times and adapted into specific domains. Thorough observation of TAM concepts reveals similarities with the malicious threat components. PEOU can be related to the opportunities and PU to motivation. Venkatesh et al. [10] reviewed the existing acceptance models and formulated the unified theory of acceptance and use of technology (UTAUT). This new model outperforms other

existing acceptance models in predicting user behavior. UTAUT model consists of the following constructs (causal relationships are depicted on Fig. 3):

- *performance expectancy (PE)*: degree to which an individual believes that using the system will help him or her to attain gains in job performance,
- *effort expectancy (EE)*: degree of ease associated with the use of system,
- *social influence (SI)*: degree to which an individual perceives that important others believe he or she should use the system,
- *facilitating conditions (FC)*: degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system.

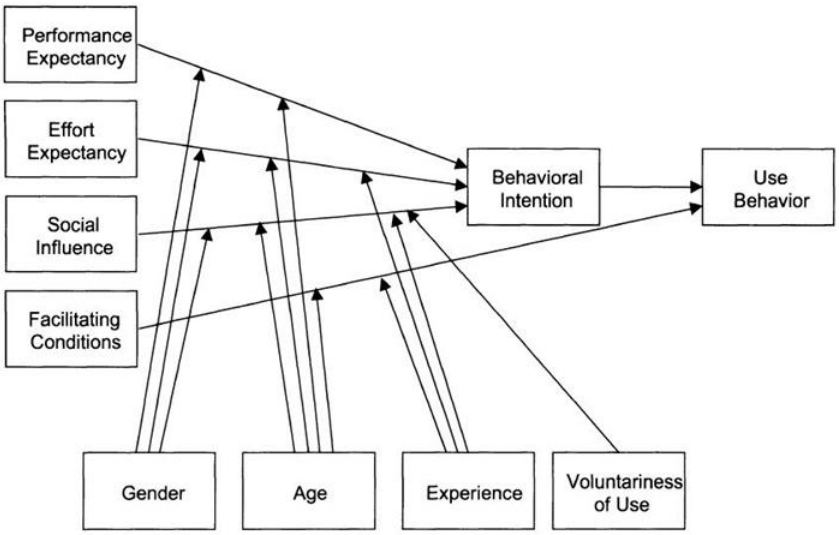


Figure 3 UTAUT model

We can find many researches using the UTAUT model in the area of eLearning systems, mobile services, process automation and some others. Up to our knowledge, no similar studies have been done in the security domain. Our idea is to adapt the UTAUT model with factors that are specific for attackers and instead of "usage" we will check the "misusage of technology". The expected result is the acceptance model, which will help us understand the mind-set of malicious threat agent on targeted system.

Another crucial building block upon which we can make predictions is the evidence of existing malicious threats. They leave trails behind their activities. Therefore, it is important to uncover the circumstances of the first occurrence of threat for the observed IS. A good approximation of the first appearance of malicious threat is the date of the first vulnerability discovery.

Thanks to some organizations (MITRE-CVE [11], NIST-NVD [12]), which take care for publication of discovered vulnerabilities of most popular products on the market, such data exist. They are publicly available and will be used in our future studies. On Fig. 4 the vulnerability discovery trail is represented for Wordpress product. We can see that the first discovery appeared in 2004 and some time before that malicious threat agent started activity. Consequently, the Wordpress became attractive for the attackers in that time.

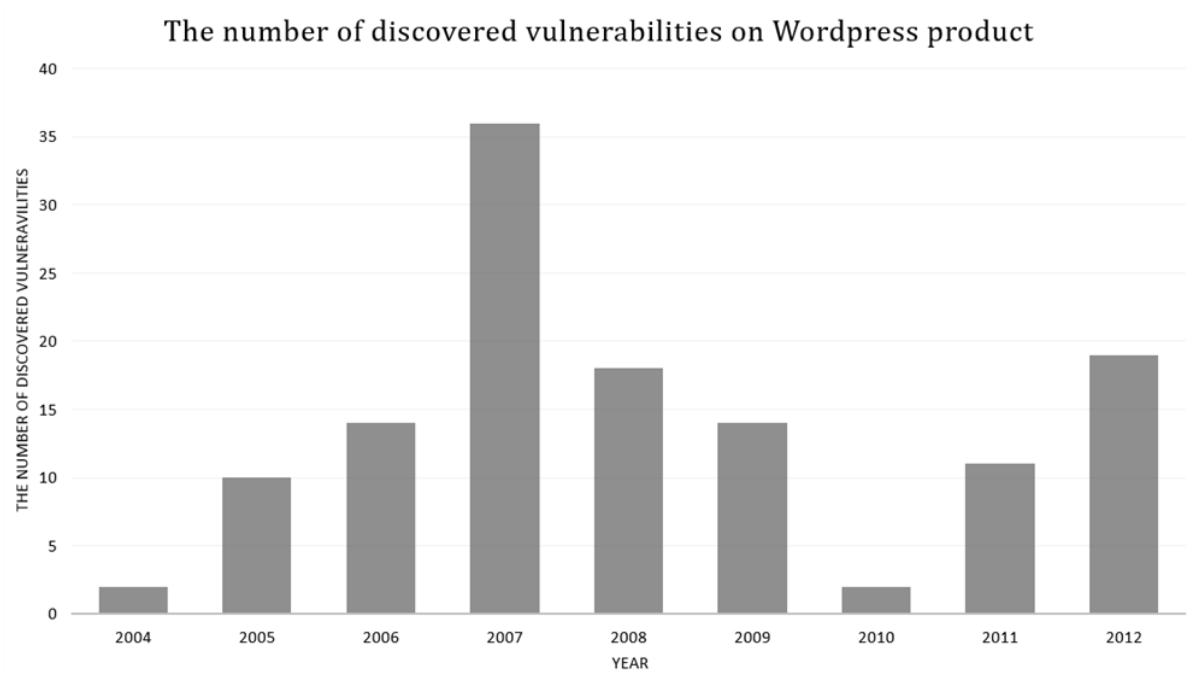


Figure 4 First vulnerability was discovered in 2004 and some time earlier threats.

4 Conclusion

Many studies have already pointed out the need to anticipate the discovery of vulnerabilities in software. However, the field of IT security is enormously intertwined with the human factor and we should pay attention on human vulnerabilities as well. Another important building block for security assessment are threats. The most complex threats for the information system are malicious threats. Behind them are always people. Therefore, we should understand the basic features about the attackers' mind-set and his or her behavior. In this paper, we presented some ideas how we could detect or predict the appearance of malicious threat. We concluded that the technology acceptance models can be useful to determine the key factors, which give birth to new malicious threats. In order to determine the real values of these key factors, we can help us with the databases of discovered vulnerabilities and other types of attacker' activity trails.

In future research we will focus on adaptation of UTAUT model for the purposes of better understanding the malicious attackers' behavior. Another research goal is to analyze the existing data about some important software products on the market in order to determine the circumstances of some key factors at the time of first vulnerability discovery.

5 Acknowledgements

Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

6 References

[1] R. J. Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems*. Wiley, 2010.

[2] O. Alhazmi and Y. Malaiya, "Prediction capabilities of vulnerability discovery models," in *Proceedings of the 17th International Symposium on Software Reliability Engineering*, 2006, no. Ref 2, pp. 343–352.

[3] I. Mann, *Hacking the Human: Social Engineering Techniques and Security Countermeasures*. Gower, 2008.

[4] S.-W. Woo, H. Joh, O. H. Alhazmi, and Y. K. Malaiya, "Modeling vulnerability discovery process in Apache and IIS HTTP servers," *Comput. Secur.*, vol. 30, no. 1, pp. 50–62, 2011.

[5] O. Alhazmi and Y. Malaiya, "Measuring and Enhancing Prediction Capabilities of Vulnerability Discovery Models for Apache and IIS HTTP Servers," *2006 17th Int. Symp. Softw. Reliab. Eng.*, pp. 343–352, 2006.

[6] S. Frei, D. Schatzmann, B. Plattner, and B. Trammell, "Modelling the Security Ecosystem - The Dynamics of (In) Security," in *Workshop on the Economics of Information Security (WEIS)*, 2009.

[7] O. H. Alhazmi, Y. K. Malaiya, and I. Ray, "Measuring, analyzing and predicting security vulnerabilities in software systems," *Comput. Secur.*, vol. 26, no. 3, pp. 219–228, 2007.

[8] S. Vidalis and A. Jones, "Analyzing Threat Agents & Their Attributes," in *Proceedings of the 5th European Conference on Information warfare and Security*, 2005, pp. 1–15.

[9] F. D. Davis, "Perceived Usefulness , Perceived Ease Of Use , And User Accep," *MIS Q.*, vol. 13, pp. 319–340, 1989.

[10] V. Venkatesh, M. Morris, G. Davis, and F. Davis, "User acceptance of information technology: Toward a unified view," *MIS Q.*, vol. 27, no. 3, pp. 425–478, 2003.

[11] MITRE, "Common Vulnerabilities and Exposures." [Online]. Available: http://cve.mitre.org/.

[12] NIST, "National Vulnerability Database." [Online]. Available: http://nvd.nist.gov/.

Constructing biological models from domain knowledge and literature

Dragana Miljkovic[1], Kristina Gruden[2], Nada Lavrač[1,3,4]
[1]Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
[2]National Institute of Biology, Ljubljana, Slovenia
[3]Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
[4]University of Nova Gorica, Vipavska 13, Nova Gorica, Slovenia
{dragana.miljkovic, nada.lavrac}@ijs.si
kristina.gruden@nib.si

Abstract: *A conventional way of manual construction of a dynamic model can be accompanied by additional steps which can speed up and enhance model construction. In our work, we have built a model of plant defence response to virus attacks based on domain knowledge and literature with few experimental datasets available for the model validation. Using solely manual knowledge engineering was not feasible due to the complexity of a plant defence system. For this reason, the laborious work of manual model creation was complemented by additional automatised and semi-automatised steps. These steps address the model structure, which was enhanced by means of natural language processing techniques, as well as model dynamics, where the parameter optimisation was guided by constraints defined by the domain experts.*

Key Words*: systems biology, plant defence, triplet extraction, natural language processing, evolutionary algorithms*

1 Introduction

In biological sciences, the growth of experimental data is not uniform for different types of biological mechanisms, hence some biological mechanisms still have few datasets available. The paper describes a novel methodology for the construction of biological models by eliciting the relevant knowledge from literature and domain experts. The methodology has been applied to build the model of defence response in plants. The developed plant defence model consists of three sub-models: salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) sub-model, since SA, JA and ET are the most important components of the defence mechanism in plants. The methodology addresses two aspects of biological model construction: the model structure and the model dynamics.

To construct the model structure, different information sources can be used. Given that most of human biological knowledge is stored in the silos of biological literature, retrieving information from the literature is necessary when building the biological models. State-of-the-art technologies enable information extraction from scientific texts in an automated way by means of text processing techniques, based on the advances in the area of natural language processing (NLP) of biology texts. NLP is used in systems biology to generate the model structures or to enhance the existing ones. The most common NLP approaches can be classified into three categories [1]: rule-based approaches, machine-learning approaches and co-occurrence-based approaches. Examples of rule-based systems include GeneWays [2], Chilibot [3], etc. There are also

combined methods, including co-occurrence-based approaches, such as Suiseki developed by Blaschke and Valencia [4]. A wide range of machine learning techniques is used for relations extraction in systems biology, like the Naive Bayes classifier [5], Support Vector Machines [6], etc. Due to the small amount of existing quantitative data, mathematical optimisation methods for parameter optimisation were recently employed in systems biology. Various local deterministic optimization techniques, like Levenberg-Marquardt algorithm [7] and stochastic approaches, like Genetic Algorithms [8] and Evolutionary Algorithms [9] are applied in systems biology.
The goal of this paper is to present the developed methodology which enables building complex biological models without or with scarce experimental data. The methodology consists of several steps, where the standard approach to the construction of dynamic models is enhanced with the following methods: a method for model structure revision by means of natural language processing techniques, a method for incremental model structure revision, and a method for automatic optimisation of model parameters guided by the expert knowledge in the form of constraints.

2 Materials and methods

An overview diagram of the iterative model construction process is shown in Fig. 1. The most relevant details of every step in this construction process are explained in the following subsections.

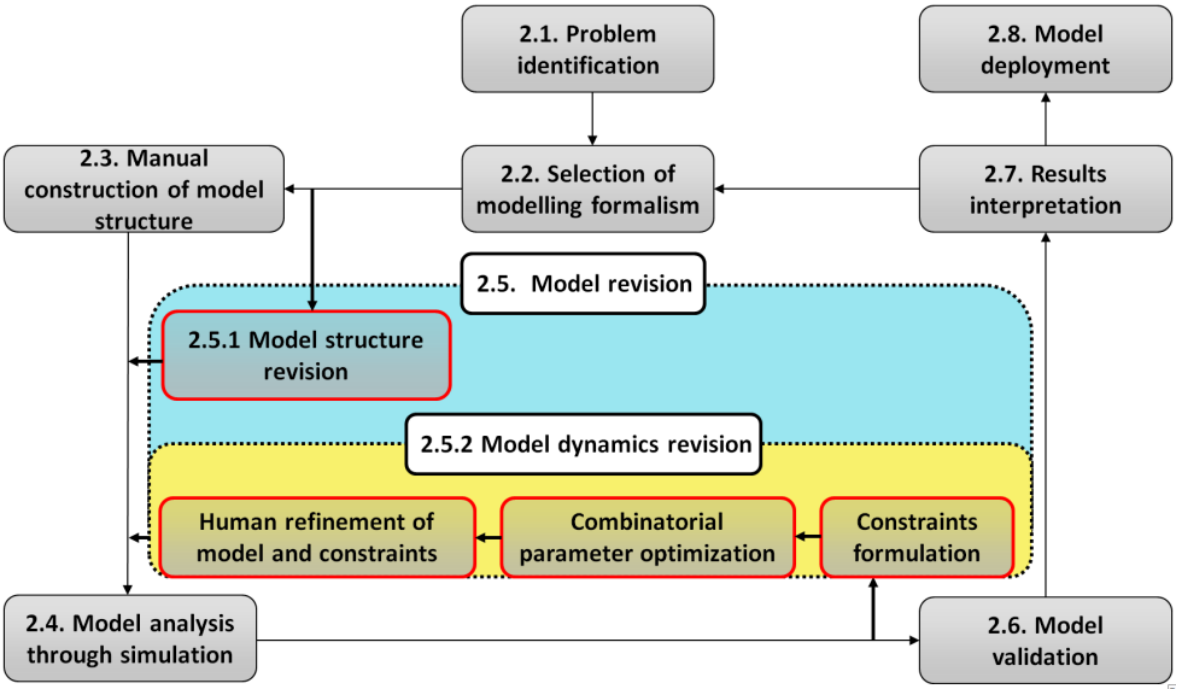


Figure 1: A schema of the developed methodology for the plant defence model construction

2.1 Problem identification

In this step we have defined the requirements for the plant defence model in collaboration with the experts from the National Institute of Biology, Ljubljana, Slovenia. The goals of model development are the following: a) better understanding of the biological mechanism on the system level and b) prediction of experimental results with the aim to detect crucial reactions in the plant defence process, predict the final defence response when some genes are silenced, and discover new connections/interactions.

2.2 Selection of modelling formalism

The decision which modelling formalism to select depends on the requirements defined by the domain experts, currently available knowledge and open issues related to the plant defence response. In plant defence research field there is few experimental data and, on the other hand, there is a lot of domain knowledge related to the modelling of the plant defence mechanism that has not yet been formalised and systematised. For these reasons, we present the model structure first in the form of a directed edge-labelled graph[1], which is a common way to present biological networks in systems biology and which is, at the same time, intuitive for the domain experts.

2.3 Manual construction of model structure

The initial plant defence model structure was constructed manually by considering knowledge from the literature, different biological databases, such as TAIR[10] and KEGG [11] and domain experts.

2.4 Model analysis through simulation

Analysis of the dynamic model behaviour was performed through iterative simulations of the manually constructed plant defence model. We selected Hybrid Functional Petri net (HFPN) formalism to simplify and simulate plant defence model, whose structure was developed initially in the graph form. The simulation was performed initially by Cell Illustrator (CI) [12], which implements the HFPN formalism. The simulator outputs time series curves of the dynamic behaviour of components of interest.

2.5 Model revision

Model revision process includes revision of model structure and its dynamics.

2.5.1 Model structure revision

The process of fusing expert knowledge and manually obtained information from the literature to build the plant defence model structure turns out to be time-consuming, non-systematic and error-prone. Therefore, we have introduced one additional step that enhances the manually built model structure: extraction of relations between biological components from the literature, using the natural language processing approach. We have developed the Bio3graph tool [13] that searches literature for the relations between the biological components and outputs a graph of triplets in the form (component1, relation, component2). These triplets were compared with the manually developed plant defence model and as a result we discovered new relations with respect to the manually constructed plant defence model. The new relations were investigated further and some of them turned out to be interesting hypotheses for further biological experiments [13]. Moreover, we have also developed an incremental version of the Bio3graph tool, which updates the network structure with new sets of triplets having an initial model as a baseline. Whenever necessary, the structures of biological networks can be quickly updated by using the incremental version of Bio3graph.

2.5.2 Model dynamics revision

This process consists of three steps:

a. Constraints formulation. The constraints are mathematical expressions defined by the domain experts. They represent the rules how the simulation output curves of certain biological components should look like. The purpose of constraints is to guide and speed up the parameter optimisation search by limiting the parameter search space.
b. Combinatorial parameter optimisation. We used differential evolution algorithm for parameter optimisation developed by Filipic and Depolli [14].

[1] Directed edge-labelled graph of the plant defence model structure represents biological components as nodes and biological reactions as vertices between them.

This algorithm performs a population-based search that optimises the problem by iteratively trying to improve a candidate solution with regard to a given measure of quality.

c. Human refinement of model and constraints. If dynamic behaviour of the curves does not fit the experts' expectations, the model structure and the constraint setting are modified.

2.6 Model validation

The validation of the simulation results is based mainly on the judgement of the domain experts to ensure that the model is close to the real-life system. However, the simulation outputs can also be validated with the validation method for non-observable systems, such as parameter sensitivity analysis.

2.7 Results interpretation

The simulation results are interpreted by the biology experts. The output curves of the simulation allow for qualitative conclusions regarding the dynamic behaviour of the model. For example, by comparing the growth of certain curves, we can determine the components that contribute the most to the plant defence response.

2.8 Model deployment

The finalised version of the model will be used at National Institute of Biology in Ljubljana (and hopefully in broader scientific community) to assist the experimental design by generating hypotheses how the plant defence response will react when particular genes are silenced.

3 Results and discussion

As a first step, the plant defence model structure is compiled manually in the form of a directed edge-labelled graph and visualised with the Biomine visualisation engine [15]. Reactions are presented as graph arcs labelled with one of three reaction types: activation (A), inhibition (I) and binding, and components as graph nodes connected with arcs to each other. We present the model structure in the form of directed edge-labelled graph in Fig. 2.A, which is a complex graph consisting of 175 nodes and 387 reactions.

Since a model structure in Fig.2A is too complex to be used directly for simulation purposes, we have simplified it and converted to be simulated with the CI software [12]. The software facilitates easy building of the network structure, has a graphical editor that has drawing capabilities and allows biologists to model different biological networks and simulate the dynamic interactions between the biological components. The final result of a model prepared for the simulation is shown in Fig.2B.

The developed model structure, shown in the form of graph in Fig.2A contains more detailed information compared to the structural model of the subsets of plant defence [16] having in total 175 components and 387 reactions.

The model structure transformed to the HFPN presentation is prepared for the analysis of model dynamics. A slightly reduced structure (see Fig.2B), compared to the directed edge-labelled graph of Fig.2A, contains in total 99 components and 68 reactions. The structure of one of the first simulation models of the plant defence response [17], containing 18 biological entities and 12 Boolean operators, is less complex than the model structure shown in Fig.2B.

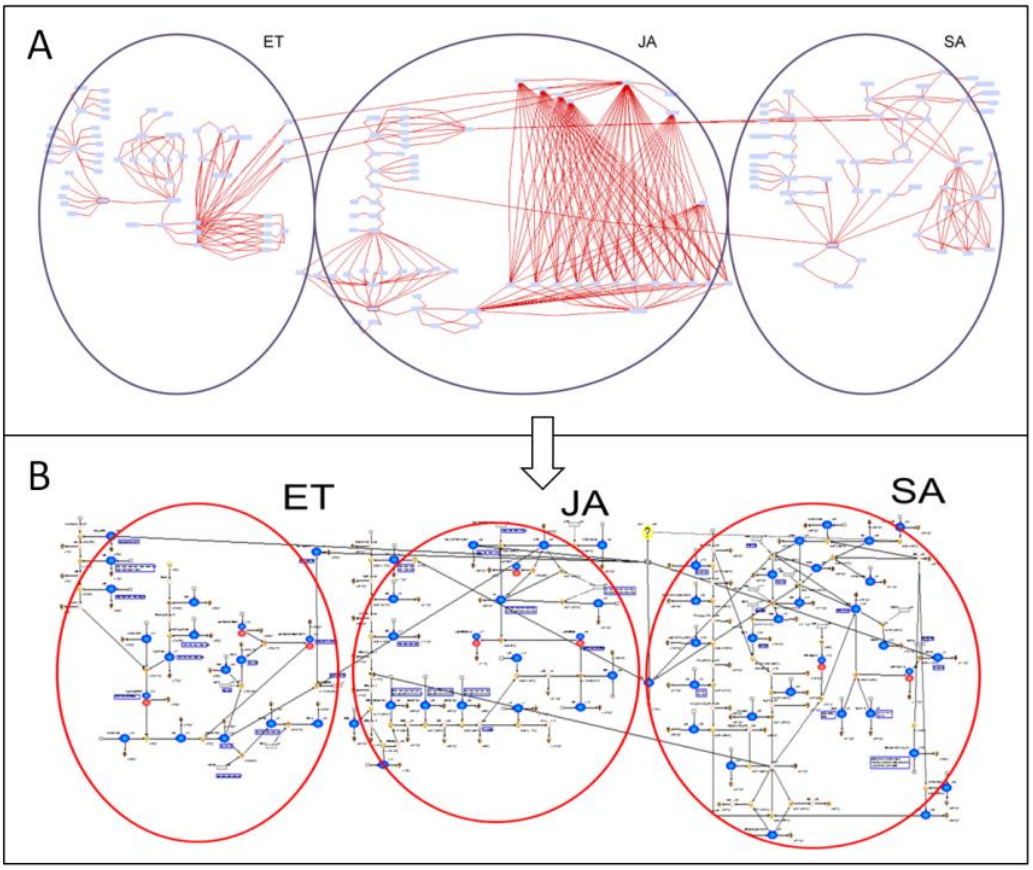


Figure 2: Plant defence model. A) Model structure in the form of a directed edge-labelled graph. B) Simplified model, developed according to the HFPN formalism for simulation purposes

8 Acknowledgements

This work was financed AD Futura Agency, Slovenian Research Agency grants P4-0165, J4-2228, J4-4165, J2-5478 and P2-0103.

9 References

[1] Cohen, K. B.; Hunter, L. Getting started in text mining. Plos Computational Biology 4, 2008.

[2] Rzhetsky, A.; Iossifov, I.; Koike, T.; Krauthammer, M.; Kra, P.; Morris, M.; Yu, H.; Duboue, P. A.; Weng, W. B.; Wilbur, W. J.; Hatzivassiloglou, V.; Friedman, C. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. Journal of Biomedical Informatics 37, 43-53, 2004.

[3] Chen, H.; Sharp, B. M. Content-rich biological network constructed by mining pubmed abstracts. Bmc Bioinformatics 5, 2004.

[4] Blaschke, C.; Valencia, A. The frame-based module of the suiseki information extraction system. Ieee Intelligent Systems 17, 14-20, 2002.

[5] Craven, M.; Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. In: Proceedings of International Conference on Intelligent Systems for Molecular Biology. 77-86, 1999.

[6] Donaldson, I.; Martin, J.; de Bruijn, B.; Wolting, C.; Lay, V.; Tuekam, B.; Zhang, S.;Baskin, B.; Bader, G.; Michalickova, K.; Pawson, T.; Hogue, C. Prebind and textomy -mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4, 2003.

[7] Levenberg, K. A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics 2, 164-168, 1944.
[8] Mitchell, M. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA, 1996.
[9] Eiben, A. E.; Smith, J. E. Introduction to Evolutionary Computing. Springer-Verlag, Berlin, 2003.
[10] Swarbreck, D.; Wilks, C.; Lamesch, P.; Berardini, T. Z.; Garcia-Hernandez, M.; Foerster, H.;Li, D.; Meyer, T.; Muller, R.; Ploetz, L.; Radenbaugh, A.; Singh, S.; Swing, V.; Tissier,C.; Zhang, P.; Huala, E. The arabidopsis information resource (tair): gene structure and function annotation. Nucleic Acids Research 36, D1009-D1014, 2008.
[11] Kanehisa, M.; Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27-30, 2000.
[12] Nagasaki, M.; Doi, A.; Matsuno, H.; S., M. Genomic object net: a platform for modeling and simulating biopathways. Applied bioinformatics 2, 181-184, 2004.
[13] Miljkovic, D.; Stare, T.; Mozetič, I.; Podpečan, V.; Petek, M.; Witek, K.; Dermastia, M.; Lavrač, N.; Gruden, K. Signalling network construction for modelling plant defence response. PLOS ONE 7, e51822-1e51822-18,2012.
[14] Filipič, B.; Depolli, M. Parallel evolutionary computation framework for single- and multiobjective optimization. Parallel Computing 217-240, 2009.
[15] Eronen, L.; Toivonnen, H. Biomine: predicting links between biological entities using network models of heterogeneous databases. BMC Bioinformatics 13, 119, 2012.
[16] Staswick, P. E.; Tiryaki, I. The oxylipin signal jasmonic acid is activated by an enzyme that conjugates it to isoleucine in arabidopsis. Plant Cell 16, 2117-2127, 2004.
[17] Genoud, T.; Santa Cruz, M. B. T.; Metraux, J. P. Numeric simulation of plant signaling networks. Plant Physiology 126, 1430-1437, 2001.

Comprehensibility of Classification Trees – Survey Design Validation

Rok Piltaver[1], **Mitja Luštrek**, **Matjaž Gams**[1]
Department of Intelligent Systems
Jozef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
{rok.piltaver, mitja.lustrek, matjaz.gams}@ijs.si

Sanda Martinčić - Ipšić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
smarti@inf.uniri.hr

Abstract: *Classifier comprehensibility is a decisive factor for practical classifier applications; however it is ill-defined and hence difficult to measure. Most algorithms use comprehensibility metrics based on classifier complexity – such as the number of leaves in a classification tree – despite evidence that they do not correspond to comprehensibility well. A classifier comprehensibility survey was therefore designed in order to derive exhaustive comprehensibility metrics better reflecting the human sense of classifier comprehensibility. This paper presents an implementation of a classification-tree comprehensibility survey based on the suggested design and empirically verifies the assumptions on which the survey design is based: the chosen respondent performance metrics measured while solving the chosen tasks can be used to indirectly but objectively measure the influence of chosen tree properties on their comprehensibility.*

Keywords*:* *classification tree, comprehensibility, understandability, survey*

1 Introduction

In data mining the comprehensibility is the ability to understand the output of an induction algorithm [11]. It is also referred to as interpretability [14] or understandability [1] and has been recognized as an important property since the early days of machine learning research [16, 18]. Although research in the last three decades is more focused on improving predictive performance of learned classifiers, comprehensibility is reported as the decisive factor when machine learning approaches are applied in industry [10]. Examples of application areas in which comprehensibility is emphasized are medicine, credit scoring, churn prediction, bioinformatics, etc.[5].

A comprehensibility metric is needed in order to compare performance of learning systems and as a (part of) heuristic function used by a learning algorithm [6, 20]. However, comprehensibility is ill-defined [10] and subjective [1, 8, 13], therefore it is difficult to measure. Instead of measuring comprehensibility directly, most algorithms

[1] The authors are also affiliated with Jozef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

use model complexity instead; e.g. number of leaves in a tree [13], number of conjunctions in a rule set [19], number of connections in an artificial neural network [9, 12]. Although model complexity is related to comprehensibility [9], empirical user studies [1] reveal that comprehensibility measures based solely on model complexity are over-simplistic and produce incomprehensible models [5].

In order to derive exhaustive comprehensibility metrics better reflecting the human sense of classifier comprehensibility, a classifier comprehensibility survey has been designed recently [17]. The user-survey design is based on the related work and follows the observation that the comprehensibility is in the eye of the beholder [15]. The ultimate goal of the survey is to obtained insights into respondents' judgments about classifier comprehensibility and to define a good comprehensibility metric.

This paper presents an implementation of classification-tree comprehensibility survey according to the suggested survey design and empirically verifies the assumptions, which the survey design is based on: the performance of respondent solving the survey tasks depends on the classification tree comprehensibility; the observed classification tree properties influence comprehensibility and the range of classification trees and questions used in the survey is broad enough to measure the influence. Finally, the survey design is based on the assumption that the objectively measured respondent performance parameters (time to answer, probability of correct answer) are related to the subjective perception of classifier comprehensibility. The list of tasks (i.e. parts of the survey) designed to measure comprehensibility includes activities: classify instance, explain classification, validate classifier, and discover new knowledge from classifier. The considered properties of classification trees are: number of leaves, depth of the tree, depth of leaves relevant to a given survey question, branching factor, and tree presentation style. In addition, the paper considers a few minor survey design choices: the order of tasks and the range and explanations of scales used to collect subjective opinions.

The paper is organized as follows. Section 2 explains the survey implementation in detail: the chosen dataset and the list of used classification trees are described in Section 2.1, each task and the set of questions for the task are described in Section 2.2. Section 3 describes the group of survey respondents, verifies the survey design assumptions and suggests improvements of the survey. Section 4 concludes the paper with discussion of the results. Appendix contains figures of all classification trees used in the survey.

2 Survey implementation

The classification-tree comprehensibility survey is implemented as an online survey in order to facilitate accurate measurements of respondent performance, remote participation, automatic checking of the correctness of answers and saving them in a database. Each question of the survey corresponds to a web page, which is dynamically generated using PHP scripts. We designed the survey around six tasks, each composed of several questions of the same type but related to different trees or parts of a tree.

The first four tasks measure the performance of respondents asked to answer questions about given classification trees. The difficulty of the questions in each task depends on comprehensibility of the classification tree – an approach advocated by some researchers [1, 8]. Each of the first four tasks is based on [3], which reports that comprehensibility is required to explain individual instance classifications, validate the classifier, and discover new knowledge. The second part of the survey measures subjective opinion about comprehensibility of classification trees rated on the scales suggested in [17].

2.1 Dataset and classification trees

All the survey questions are related to the Zoo domain from the UCI Machine Learning Repository [2]. The domain was chosen because it meets all the requirements stated in the survey design: it is familiar and interesting to the general and heterogeneous population of respondents, but still broad and rich enough to enable learning a range of classifiers with various properties. The Zoo domain requires only elementary knowledge about animal properties expressed with 16 (mostly binary) attributes: hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs (numeric), tail, domestic, and catsize. The attribute *animal name* from the original dataset is not used in the survey because it is a unique identifier of an instance. The seven classes given as numeric attribute in the original Zoo domain are referred to using the following names instead: mammals (41 instances), birds (20), fish (13), mollusc (10), insect (8), reptile (5), and amphibian (4).

The classification trees used in the survey (Figures 3-20) are learned and visualized using the Orange tool [4] as suggested by [17]. The basic tree shown in Figure 4 is learned using the Classification Tree widget in Orange with the following parameters: gini index as the attribute selection criterion, pruning with m-estimate where $m = 2$, and minimum of 4 instances in a leaf. It contains 7 leaves, 6 binary attributes (resulting in branching factor 2), and depth 5. Choosing other attribute selection criterion would only change the order of the same set of attributes in the tree.

The survey is based on trees with three different sizes: small trees with 3 or 4 leaves, the basic tree with 7 leaves, and big trees with 9 or 10 leaves. The big trees (Figure 5) were learned on the original dataset using modified pruning parameters. The two small trees are shown in Figure 3. They were learned using a reduced set of attributes because such trees are more natural and comprehensible then the pruned versions of the basic tree; which is caused by the uneven class distribution in the dataset. In this way an unnatural classifiers – a possible survey design error [1] – was avoided. The sets of attributes used to learn the small trees were chosen so that the leaves of the learned trees correspond to clusters obtained using hierarchical clustering (by Euclidian distance and complete linkage criterion) on the original dataset. As a result, each leaf in the learned trees contains animals from a single class or classes of animals with similar properties.

In addition, the survey analyses influence of tree branching factor on comprehensibility, therefore the trees described above were modified to obtain trees with branching factor 3 (Figures 6-8) and 4 (Figures 9-11). This was achieved by adding new aggregate attributes, which were computed as natural combinations of original attributes selected so that the structure of the trees with higher branching factor is as similar as possible to the structure of the trees with branching factor 2. Note that two version of the tree with 9 leaves (Figure 8) and branching factor 3 were obtained, yet they differ in the maximal depth of the tree.

In addition to the trees described above, their modified representations (Figures 12-17) are used in the last task as discussed in the description of the *compare* task. The trees shown in Figures 18-20 are used only in the *discover* task and were obtained using the Orange Interactive Tree Builder as discussed in the paragraph about the task.

2.2 Tasks and questions

Each of the tasks starts with an instructions page. It includes an explanation of the task on an example: figures and explanations showing how to solve the task step by step. The instruction page includes an example question in exactly the same format as the questions in the task but on a different domain. The respondents are allowed to start answering the questions only after correctly answering the test question. The test

question was added to the survey because significantly longer times of answering the first question compared to the subsequent questions were observed in the initial testing.

The first task – **classify** – asks a respondent to classify an instance according to a given classification tree – the same task type was used in [1, 8]. When a webpage with a survey question (Figure 1) is opened, only the instructions and footer of the page are shown. The instruction for the first task says: "*Classify the instance described in the table on the left using the classification tree on the right.*" The footer contains respondent's name, name of the task and current question number, link to a help page, and support e-mail address. After reading the instructions, the respondent clicks the "*Start solving*" button. This calls a JavaScript function that starts the timer and displays the question data and the answer form. The question data consists of a table with ten alphabetically sorted attribute-value pairs shown on the left and an image of a classification tree in SVG format shown on the right (Figure 1). The answer form is displayed below the question data; the label says: "*The instance belongs to class*:" and is followed by a drop-down menu offering the names of the seven classes as an answer to the question. Whenever the respondent changes a value of an answer form field, the time and action type are recorded. When the respondent clicks the "*Submit answer*" button, the answer fields are disabled, the timer is stopped and the time needed to answer the question is calculated.

In addition, the respondent is asked to give the subjective judgment of questions difficulty on the scale with five levels. Each label of the scale is accompanied with an explanation in order to prevent variation in subjective interpretations of the scale:

- Very easy – I answered without any problems in less than 5 seconds.
- Easy – I found the answer quite quickly and without major problems.
- Medium.
- Difficult – I had to think hard and am not sure if I answered correctly.
- Very difficult – Despite thinking very hard my answer is likely to be wrong.

After rating the question's difficulty, the respondent clicks the "*Next question*" button, which calls a JavaScript function that assigns the calculated performance values to the hidden form fields in order to pass them to the PHP script that stores the data in the database and displays the next question. One question per each leaf depth was asked for each tree shown in Figures 1-11, which amounts to 30 questions. The number of questions in other tasks was reduced because the first group of respondents reported that the number of questions should not be any higher in order to prevent them from becoming tired or bored while answering the survey.

The second task – **explain** – asks a respondent to answer which attributes' values must be changed or retained in order for the tree to classify the given instance into another class. This corresponds to explaining an individual instance classification. For example, which habits (values of attributes) would a patient with high probability of getting cancer (class) have to change in order to stay healthy? The web pages for questions in the second task are similar to the ones in the first task with the following differences. The instruction for the task says for example: "*What should be the values of the attributes for the example (listed in the table on the left) that is classified as fish so that the classification tree (shown on the right) would classify it as amphibian? If the value of an attribute is not relevant for the classification into the new class, maintain the default choice (i.e. "irrelevant"); otherwise change it to "unchanged" if the value must remain as it is, or "different" if it needs to be changed.*" The table of attribute-value pairs includes an additional column named "*new value*" with drop-down menus with three choices: irrelevant (default value), different, and unchanged.

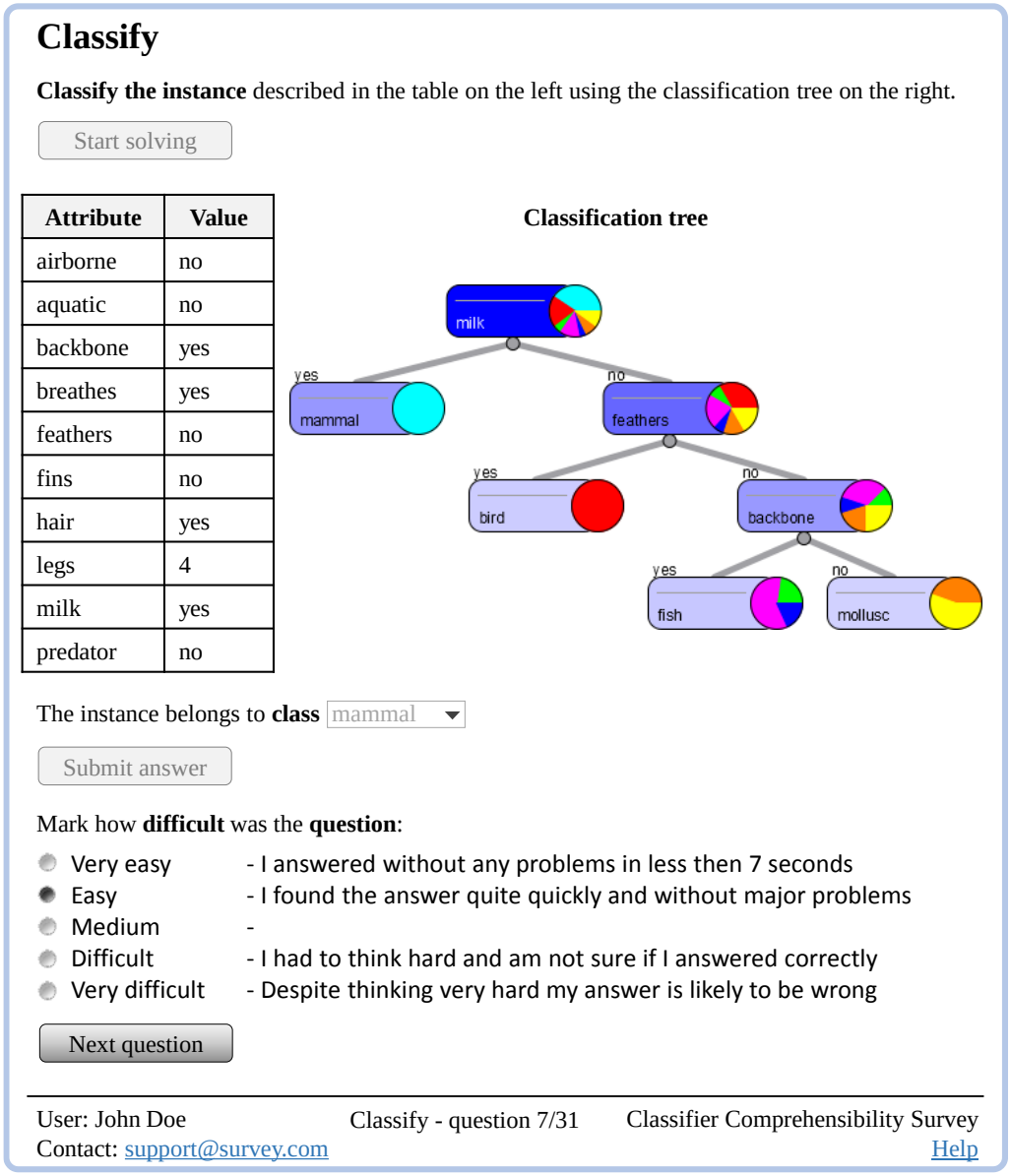


Figure 1: Web page with an example of question from the classify task.

The third task – **validate** – asks a respondent to check whether a statement about the domain is confirmed or rejected by the presented tree – this corresponds to validating a part of the classification tree. Similar questions were also asked in [8]. The question web pages for the third task are similar to the ones in the first task with the following differences. The instruction for the task says: "*Does the classification tree agree with the statement below it?*" The answer form is replaced with a statement, e.g.: "*For animals from class reptile it holds that aquatic = yes and feathers = no*", followed by a drop-down menu with two possible answers: yes or no. Each statement is composed of two conditions (for two different attributes), except when knowledge about animal class belonging to a leaf at depth one is verified, e.g. questions about mammals for the trees in Figure 3. The number of conditions to be validated is limited to two in order to observe the relative difficulty of validating domain knowledge corresponding to leaves at various depths regardless of the number of attributes relevant for classification of the instances belonging to those leaves – this is already considered in the classify and explain tasks.

The fourth task – **discover** – asks the respondent to find a property (attribute-value pair) that is unusual for instances from one class, which corresponds to discovering new knowledge from the classification tree. Rather than rediscovering known relations

between attributes and classes, the questions ask to find an unusual property for a class of animals – a common property of outliers, e.g. it is unusual for a mammal to lay eggs. Therefore special trees offering the information about the outliers were constructed. The outliers in the dataset were first identified using the Outliers widget in Orange. After that the trees that misclassify the outliers were constructed using the Interactive Tree Builder widget in Orange. Some parts of the tree were selected manually in order to place the attribute that splits the outliers from the normal instances (the attribute expected as the answer) at a desired depth in the tree. The remaining parts of the tree were built using the automatic tree learning widget function. In this way the four trees shown in Figures 18-20 were constructed. In contrast with the trees used in other tasks, each of their nodes includes a number of instances of a given class belonging to the node – this is used by respondents to identify the common and rare properties of animals belonging in to a class. A question for each node containing outlier instances observable in the trees shown in Figures 18-20 was asked amounting to 8 questions.

The fifth task – **rate** – requests the user to give the subjective opinion about the classification trees on a scale with five levels:

- Very easy to comprehend – I can use the knowledge represented by the classification tree as soon as I read it for the first time; I can easily remember it; I can explain it to another person without looking at the figure.
- Easy to comprehend – I can use the knowledge represented by the classification tree after studying it for some time; I can remember it with some effort; I can explain it to another person without looking at the figure, but I am not sure that the person will fully understand it.
- Comprehensible – Without long study I can use most of the knowledge represented by the classification tree, but need the figure for some details; it would be difficult to remember it; I can explain it to another person if I see the figure while the other person does not see it, but that person is unlikely to fully understand it.
- Difficult to comprehend – Without long study I can use only some of the knowledge represented by the classification tree, and need the figure for the rest; It would be very difficult to remember it; I can explain the outline to another person if I see the figure while the other person does not see it.
- Very difficult to comprehend – I need the figure to use the knowledge represented by the classification tree; it would be extremely difficult to remember it; I can explain it to another person only if we both see the figure.

Each label of the scale is accompanied with an explanation in order to prevent variation in subjective interpretations of the scale. The web pages of questions in the fourth task are again similar to the ones in the first task with some differences. Namely, the instruction for the task says: "*How comprehensible is the tree shown below?*" Additionally, the answer form is replaced with the table containing the comprehensibility scale and a radio button for each of the comprehensibility levels. No attribute-value table is shown in this task. The respondents were asked to rate the comprehensibility of each tree 12 trees shown in Figures 3-11.

Task six – **compare** – asks the respondents to rate which of the two classification trees shown side by side is more comprehensible on the scale with four levels. The instructions say: "*The following question type measures the subjective opinion about the tree comprehensibility; there are no correct and wrong answers and time needed to answer each question is not important at all. Compare the classification trees in the pictures and choose the answer that best fits your opinion*." Clicking on a tree opens a new window showing full-screen picture of the selected tree and a back button. This is needed because text in nodes of some of the bigger trees becomes difficult to read when

the trees are scaled to half of the screen width. The answer is given by clicking a radio button in front of one of the following answers:

- The tree on the left is much more comprehensible: answering questions similar to the ones in this survey about the other tree would be significantly more difficult and would definitely take more time.
- The tree on the left is more comprehensible: answering questions similar to the ones in this survey about the other tree would be more difficult or take more time.
- The tree on the left is slightly more comprehensible: although one tree is slightly more comprehensible, answering questions similar to the ones in this survey about the other tree would not be more difficult.
- The trees are equally comprehensible: I don't notice any difference in comprehensibility between the two trees.

The three answers preferring the tree on the right are also offered, because the trees in each question are positioned to the left or the right side randomly. One of the trees in this task is already used in the previous five tasks – serving as a known frame of reference – while the other one is a previously unseen tree with the same content but represented in different style. Figure 12 shows a version of the tree without pie charts that represent learning dataset class distribution in the nodes of the tree. The pie-charts enable the user to find leaves with the same class (same prevalent colour of the pie-chart) quickly and provide additional information in easy-to-read graphical form. Figure 13 shows a version of a tree with meaningless attribute and attribute value names, which makes it more difficult to comprehend the tree, because domain knowledge cannot be used and because remembering numeric attribute names and values is more difficult than remembering known and meaningful semantic names. Figure 14 shows a version of a tree as drawn by Weka data mining program [7]; it is compared to the Orange representation of the same trees. Figure 15 shows a version of a tree in plain text output as obtained from Weka. The layout of the tree nodes in the plain text version is in general more difficult to read than the Orange output. Figure 16 shows a version of the tree with different layout of the tree branches than the rest of the trees; the other trees used in the survey place shallow subtrees to the left and deep subtrees to the right, which corresponds to the natural reading order in the western culture – from top to bottom and from left to right. The subtrees in the tree in Figure 16, on the other hand, are scrambled regardless of their depth. The tree in Figure 17 uses the default Orange format, but is learned using a subset of attributes with less evident relation with the class, e.g. without milk (which is true only for mammals) and feathers (which is true only for birds) attributes. Nevertheless, the classification accuracy of the tree is similar to classification accuracy of the tree with the same number of leaves learned on the entire dataset. In addition, three questions comparing comprehensibility of the trees with the same representation were asked: comparing the trees with the same number of nodes but different branching factor (Figure 5 vs. bottom tree in Figure 8), comparing the trees with the same branching factor but different number of nodes (Figure 5 vs. the right tree in Figure 3), and comparing the trees with the same branching factor and number of nodes but different structure resulting in different depth of the tree (trees in Figure 8).

3 Survey design verification

In order to verify the survey design that is based on the related work, the survey was filled in by group of 18 students and teaching assistants from Department of Informatics – University of Rijeka and the collected answers were analysed.

Figure 2 shows the range of trees used in the survey according to the comprehensibility rated by the respondents (the rate task). The percentages of answers

for each of the five comprehensibility levels and for each tree used in tasks classify, explain, validate, and compare are shown. The average rating for the most comprehensible tree on a scale from 1 (very easy to comprehend) to 5 (very difficult to comprehend) is 1.44 and the average rating for the least comprehensible tree is 3.83. The tree comprehensibility is well spread from very easy to medium or difficult to comprehend trees; therefore the range of trees is validated as appropriate. The survey implementation lacks a tree that is very difficult to comprehend; however, it is not possible to learn a trees that is very difficult to comprehend in the Zoo domain. Using a more complex domain would allow the construction of such trees, but might not be able to provide trees that are very easy to comprehend and not over-simplistic at the same time. In addition, classification trees are known as one of the most comprehensible classification models and might not be very difficult to comprehend even in the demanding domains. Figure 2 shows that respondent's ratings of the tree comprehensibility agree well, therefore the scale is confirmed as appropriate.

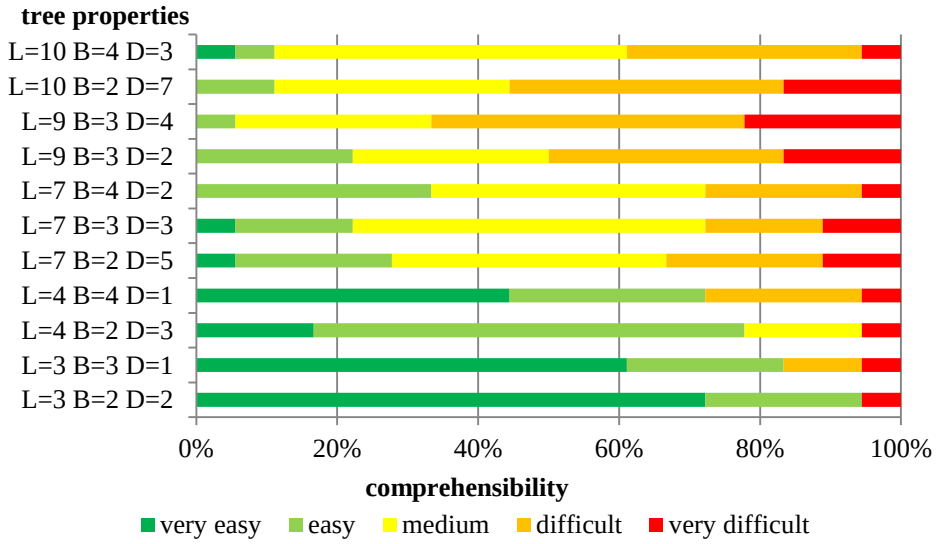


Figure 2: Subjective opinion about comprehensibility of various classification trees (L is the number of leaves in a tree, B is the branching factor, and D is the depth of a tree).

The data about the four tasks (classify, explain, verify, and discover) for which the respondent performance was measured is in Table 1. The first task (classify) was confirmed as the easiest: 98.1 % answers in the task were correct and the difficulty of individual questions in the classify task were rated from minimum 1.33 (very easy) to maximum 2.06 (easy) with an average of 1.63 on the scale from 1 (very easy) to 5 (very difficult). The fourth task (discover) was confirmed as the most difficult: 62.5 % of the questions were answered correctly and the difficulty of questions was rated from 2.00 (easy) to 2.82 (medium). The difficulty of the second and third tasks (explain and verify) are positioned between the difficulty of classify and discover tasks, as expected and incorporated into the survey design.

According to percent of the correct answers, the explain task is easier; however according to the rated question difficulty the verify task is slightly easier. The times of solving the questions from the explain tasks are longer then in verify and classify tasks, however this is partially caused by more mouse clicks required to answer a question. Based on the additional statistics it was estimated that the respondents needed about 1 to 2.5 seconds to select an answer from a drop-down menu. If this time is subtracted from the measured total time to answer a question, the time needed to reason about the most difficult question in the explain task (6 drop-down menu selections) is between 35 and 45 seconds (17 to 21 seconds per question on average). This suggests that difficulty of

the explain task is similar to the difficulty of the verify task in terms of the time needed to answer the question as well. The above observations verify that the tasks are ordered from the easiest to the most difficult ones (the order of explain and verify tasks could be switched as well).

The range of questions in each task according to their difficulty is broad: there are substantial differences between minimum and maximum time needed to solve a question in each task (see Table 1). The difficulty of questions ranges for about one level on the difficulty scale within each task. The range is even greater if individual respondents' rates are considered instead of the average rate over all the respondents.

Table 1: Overall statistics for the tasks and questions used in the survey.

task	time-difficulty correlation	correct-difficulty correlation	correct answers (%)	question time (ms) min	avr	max	question difficulty min	avr	max
classify	0.772	0.094	98.1	8.6	16.7	31.4	1.33	1.63	2.06
explain	0.957	-0.432	92.0	8.4	24.7	50.8	1.50	2.02	2.61
verify	0.720	-0.658	96.4	7.6	14.9	22.1	1.50	1.95	2.33
discover	0.901	0.548	62.5	12.7	28.6	44.6	2.00	2.53	2.82

The correlation between the rated difficulty of question and the two objective measure of question difficulty – the time needed to answer a question and the percent of correct answers – were calculated in order to verify whether they can be used to estimate the question difficulty and the tree comprehensibility objectively. The time to answer a question (averaged over all respondents that answered correctly) is clearly correlated with the rated question difficulty: the correlation ranges from 0.720 to 0.957 across the four tasks. The correlation of the percent of correctly answered questions and the rated question difficulty is almost zero for the classify task, because almost all questions were answered correctly; the task is so easy that the percent of correct answers does not change over the trees used in the survey. For the explain and the verify tasks the correlation is -0.432 and -0.658 respectively – this means that respondents correctly answered fewer questions that they rated as more difficult compared to the questions rated as easier. Interestingly, the correlation is positive in the discover task in which only 62.5 % of questions were answered correctly. If the respondents who did not know how to solve the task (rated all the questions as very difficult and answered most of them incorrectly) and the respondents who did not understand the task (rated questions as very easy or easy but answered almost all of them incorrectly) are removed, the correlation drops to 0.135. The few remaining respondents are mostly data mining experts, therefore they rated all the questions in the discover task as easy or of medium difficulty. This suggests that the survey should be slightly modified in order to be more understandable for the non-expert population.

Correlations of tree properties and the respondents' performance as well as their subjective opinions about comprehensibility of various trees is analysed in order to validate the importance and interestingness of the observed tree parameters. Correlation between the number of leaves in a tree and the rated comprehensibility of the tree is 0.951, the correlation of the number of leaves in a tree and the rated question difficulty ranges from 0.452 to 0.802 across the first four tasks, and the correlation between the number of leaves in a tree and the time needed to correctly answer the question ranges from 0.418 to 0.663 across the first four tasks. In addition the respondents rated a tree with 4 leaves as more comprehensible then tree with 10 leaves with rate 2.89 on a scale from 1 (same comprehensibility) to 4 (much more comprehensible). This supports the hypothesis that increasing number of leaves in a tree decreases its comprehensibility. Similar conclusion can be drawn for the depth of the tree; however the correlations are

lower for this parameter. Interestingly, the correlations with the branching factor are negative and close to zero, although trees with high branching factor were expected to be less comprehensible. There are two possible reasons for this. First, the difference between branching factor 2 and 4 used in the survey is not very important, therefore trees with higher branching factor should be considered in the survey. Second, increasing the branching factor does not decrease the comprehensibility because it decreases the depth of the tree and the leaves with the question answers at the same time. This explanation is supported by [19], which advocates deep model theories, which correspond to the aggregate attributes used in the survey to produce the trees with branching factor higher than two. In any case, the influence of the branching factor on the comprehensibility of classification trees should be investigated further. Another interesting conclusion of the analysis is that the depth of the question is even more correlated to respondents' performance than the number of leaves in the tree: the correlations range from 0.606 to 0.943. Therefore more emphasis should be given to this parameter in further studies. The related work [1, 8] considers only the comprehensibility of a classification tree as a whole instead of emphasizing the parts of the tree that classify more instances, are used more often, or contain instances whose correct classification and explanation is more important for the user. Finally, the presentation of classification tree clearly influences their comprehensibility and should be studied in detail. As expected, the respondents rated a simple text representation of a tree much more difficult to comprehend than a tree output produced by Orange (rate 3.72 on scale from 1 to 4). The second most important presentation property turns out to be the meaningfulness of class names and attribute values and names (rate 2.44). The arrangement of branches and Weka output versus the Orange output (rates 1.89 and 1.83) influences the comprehensibility ratings as well.

4 Conclusion

The paper presents an implementation of tree comprehensibility survey according to classifier comprehensibility survey design [17]. The analysis of the data obtained form 18 respondents (mainly CS students and some DM experts) supports that the design of tasks (classify instance, explain classification, validate classifier, and discover new knowledge from classifier) and questions is appropriate to measure the classification-tree comprehensibility. The investigated properties of classification trees (quantitative measures: number of leaves, depth of the tree and depth of relevant leaves, branching factor; and tree presentation style) show to be relevant for tree comprehensibility evaluation as well. Furthermore, the obtained results supports that objectively measured respondent performance parameters (time to answer, probability of correct answer) can be used to estimate the comprehensibility of classification trees. In addition, a few minor survey design choices are confirmed to be correct. The data provided by the 18 respondents suffices to validate the survey design choices; however, the implementation of the survey with few minor improvements (e.g. clearer instructions in the discover task, additional questions related to tree presentation in the compare task, questions related to additional trees: less comprehensible and with branching factor more than 4) should be performed with more respondents in order to obtain enough data to perform statistical analysis about the influence of various classification tree parameters on tree comprehensibility. Finally, in the future work, data mining methods will be employed in comprehensibility evaluation task, since they might prove useful in explaining the interplay of various parameters, and deriving a formal model of classification-tree comprehensibility.

9 References

[1] Allahyari, H.; Lavesson, N. User-oriented Assessment of Classification Model Understandability, 11th Scandinavian Conference on AI, pages 11-19, 2011.

[2] Bache, K.; Lichman, M. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, downloaded: July 2014.

[3] Craven, M. W.; Shavlik, J. W. Extracting Comprehensible Concept Representations from Trained Neural Networks. Working Notes on the IJCAI'95 WS on Comprehensibility in ML, pages 61-75, 1995.

[4] Demšar, J.; Curk,T.; Erjavec, A. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14(Aug):2349−2353, 2013.

[5] Freitas, A. A. Comprehensible classification models – a position paper. ACM SIGKDD Explorations, 15 (1): 1-10, 2013.

[6] Giraud-Carrier, C. Beyond predictive accuracy: what? In Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation, pages 78-85, 1998.

[7] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1):10-18 2009.

[8] Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J.; Baesens, B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems, 51(1):141-154, 2011.

[9] Jin, Y. Pareto-Based Multiobjective Machine Learning – An Overview and Case Studies, IEE transactions on systems, man, and cybernetics-part c: applications and reviews, 28(3):397-415, 2008.

[10] Kodratoff, Y. The comprehensibility manifesto, KDD Nuggets, 94:9, 1994.

[11] Kohavi, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. Proceedings of the 2nd Int. Conf. on KD and DM, pages. 202-207, 1996.

[12] Liu, G. P.; Kadirkamanathan, V. Learning with multi-objective criteria. In Proceedings of IEE Conference on Artificial Neural Networks, pages 53-58, 1995.

[13] Maimon, O.; Rokach, L. Data Mining and Knowledge Discovery Handbook. Springer, 2005.

[14] Maimon, O.; Rokach, L. Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications, World Scientific Publishing Company, 2005.

[15] Martens, D.; Vanthienen, J.; Verbeke, W.; Baesens, B. Performance of classification models from a user perspective. Decision Support Systems, 51(4):782-793, 2011.

[16] Michalski, R. A theory and methodology of inductive learning. Artificial Intelligence 20:111-161, 1983.

[17] Piltaver, R.; Luštrek, M.; Gams, M.; Martinčić - Ipšić, S. Comprehensibility of classification trees – survey design. In Proceedings of 17th International multiconference Information Society, pages 70-73, Ljubljana, Slovenia, 2014.

[18] Quinlan, J.R. Some elements of machine learning. In Proceedings of 16th International Conference on ML (ICML-99), pages 523-525, Bled, Slovenia, 1999.

[19] Sommer, E. An approach to quantifying the quality of induced theories. In Proceedings of the IJCAI Workshop on ML and Comprehensibility, 1995.

[20] Zhou, Z. H. Comprehensibility of data mining algorithms. Encyclopaedia of Data Warehousing and Mining. Idea Group Publishing, USA, 2005, pages 190-195.

Appendix: classification trees used in the survey

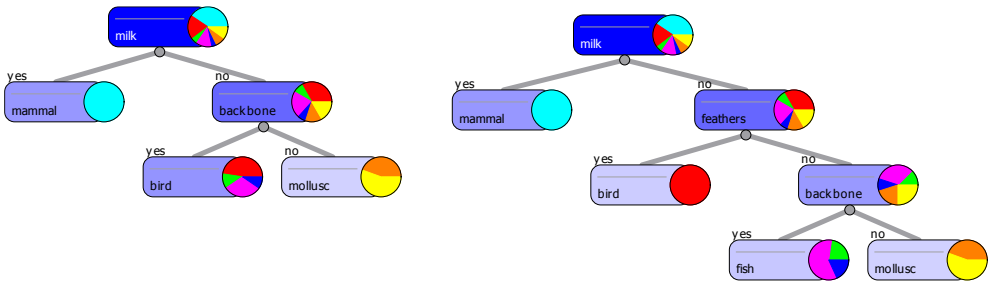


Figure 3: trees with 3 or 4 leaves and branching factor 2.

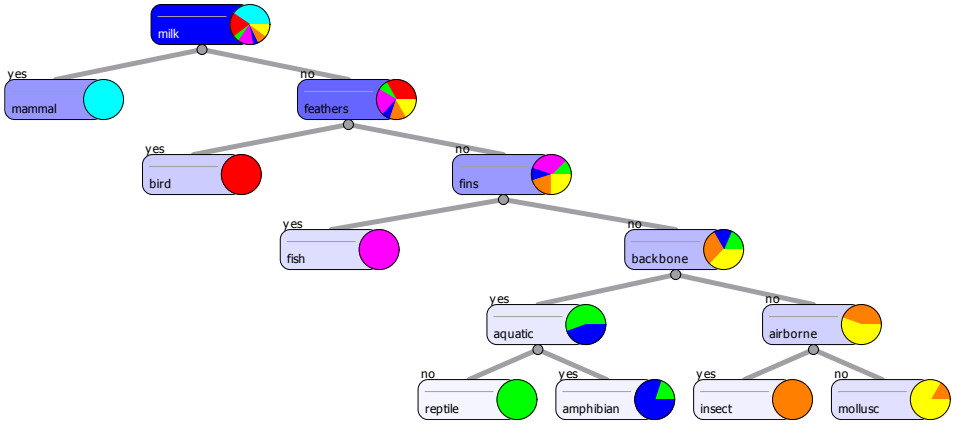


Figure 4: tree with 7 leaves and branching factor 2 – learned using the default parameters of the Classification Tree widget in Orange.

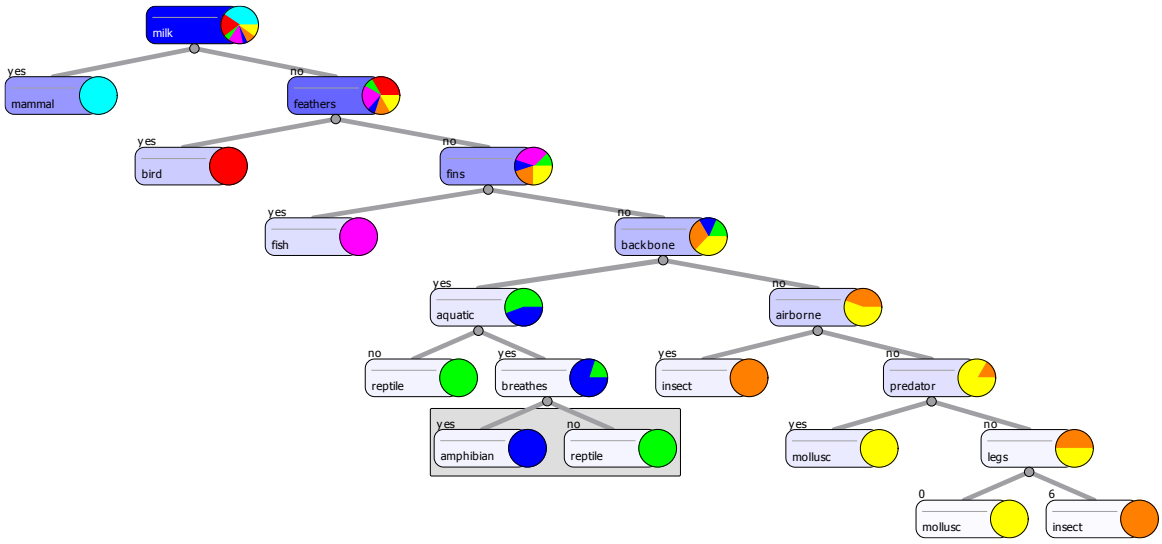


Figure 5: trees with 9 (without the subtree marked with grey rectangle) or 10 leaves and branching factor 2 – unpruned version of the tree in Figure 4.

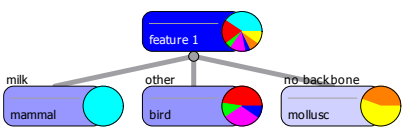


Figure 6: version of the left tree in Figure 3 with branching factor 3.

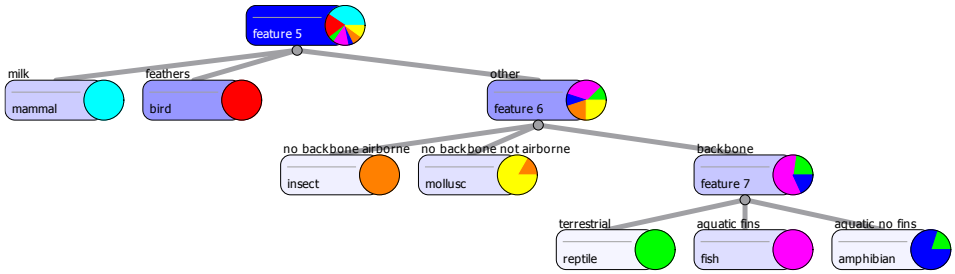


Figure 7: version of the tree in Figure 4 with branching factor 3.

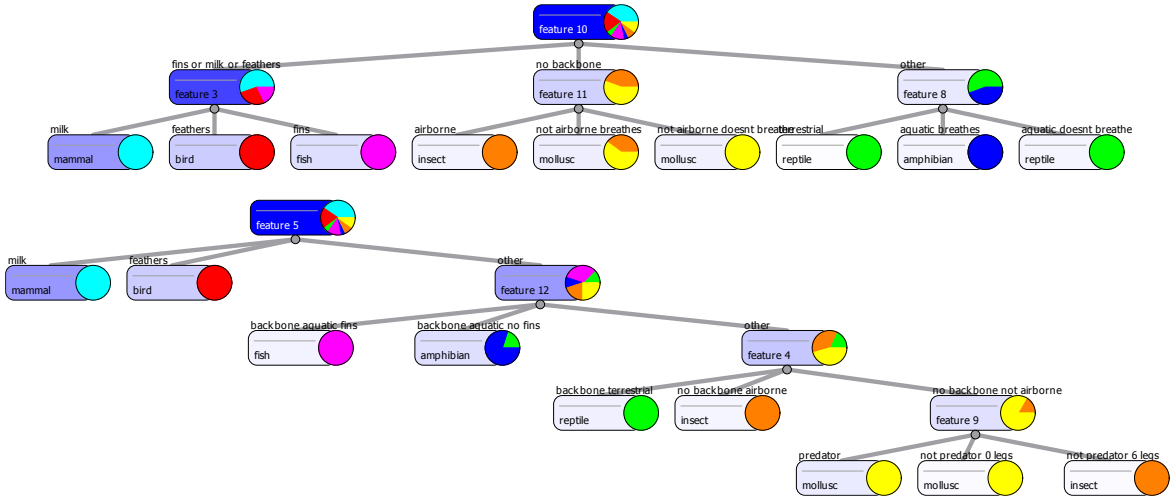


Figure 8: two version of the tree in Figure 5 with branching factor 3.

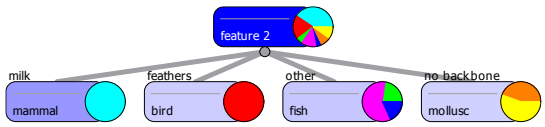


Figure 9: version of the right tree in Figure 3 with branching factor 4.

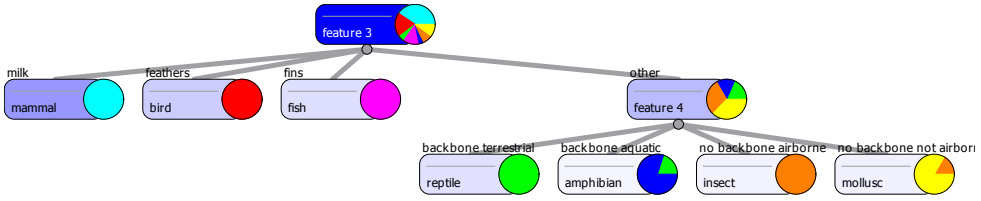


Figure 10: version of the tree in Figure 4 with branching factor 4.

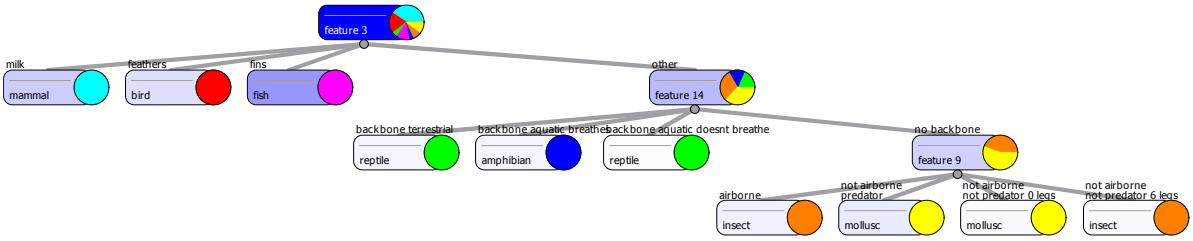


Figure 11: version of the tree in Figure 5 with branching factor 4.

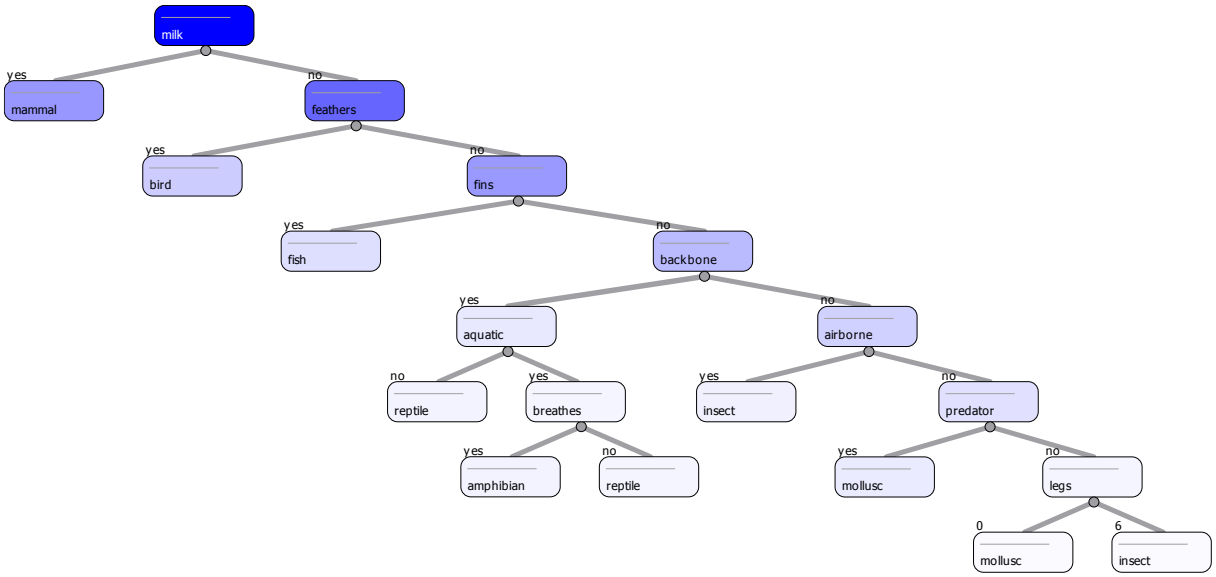


Figure 12: A version of the tree shown in Figure 5 without pie charts.

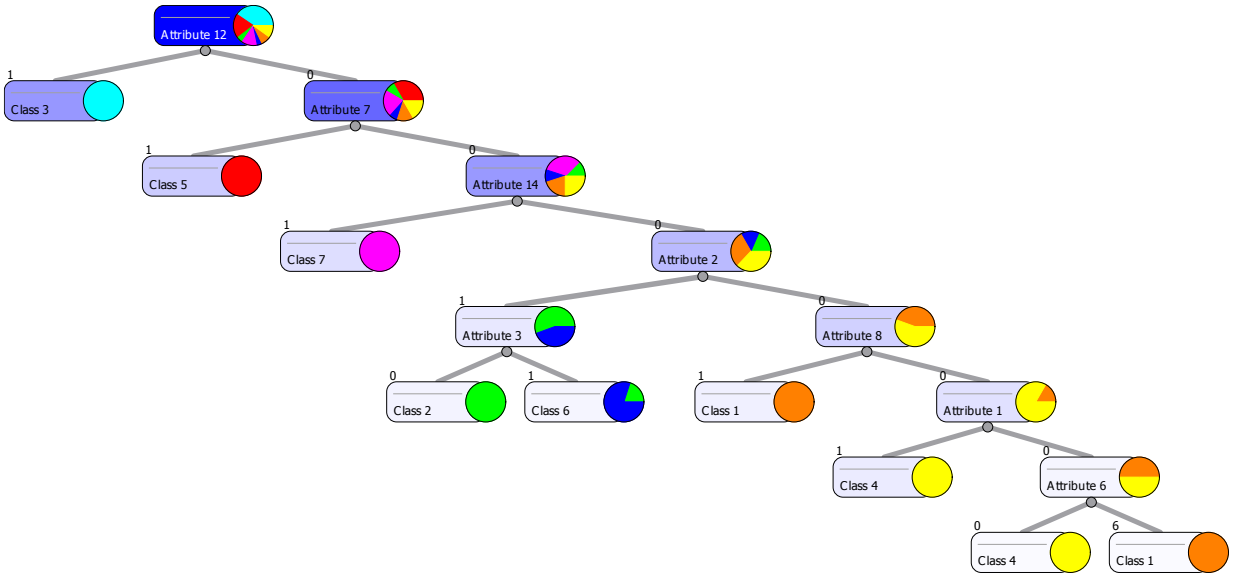


Figure 13: A version of the tree shown in Figure 5 with meaningless names of attributes and attribute values.

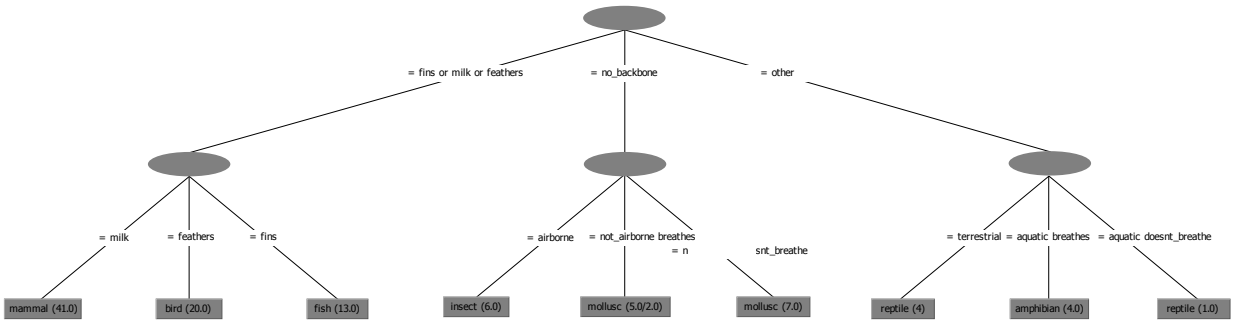


Figure 14: A version of the top tree shown in Figure 8 in Weka representation.

```
milk = no
|   feathers = no
|   |   fins = no
|   |   |   backbone = yes
|   |   |   |   aquatic = no: reptile (4.0)
|   |   |   |   aquatic = yes: amphibian (5.0/
|   |   |   backbone = no
|   |   |   |   airborne = no: mollusc (12.0/2
|   |   |   |   airborne = yes: insect (6.0)
|   |   fins = yes: fish (13.0)
|   feathers = yes: bird (20.0)
milk = yes: mammal (41.0)
```

Figure 15: A version of the tree shown in Figure 4 in plain-text format.

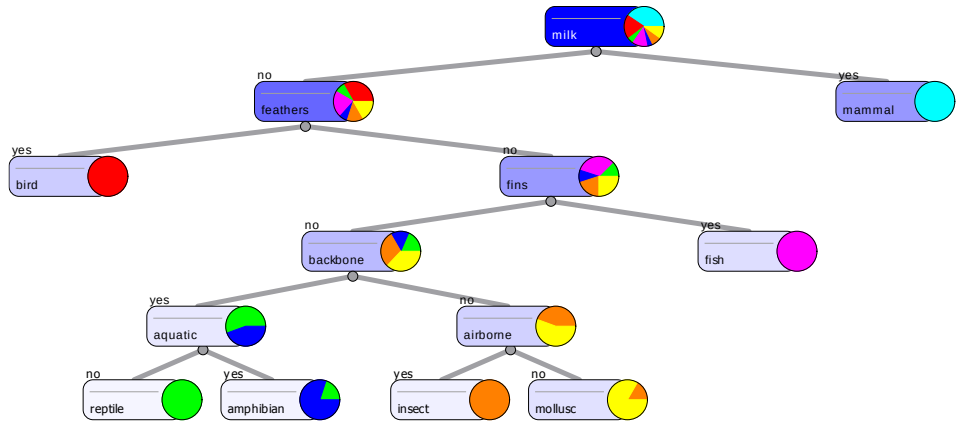


Figure 16: A version of the tree shown in Figure 4 with different layout of tree branches.

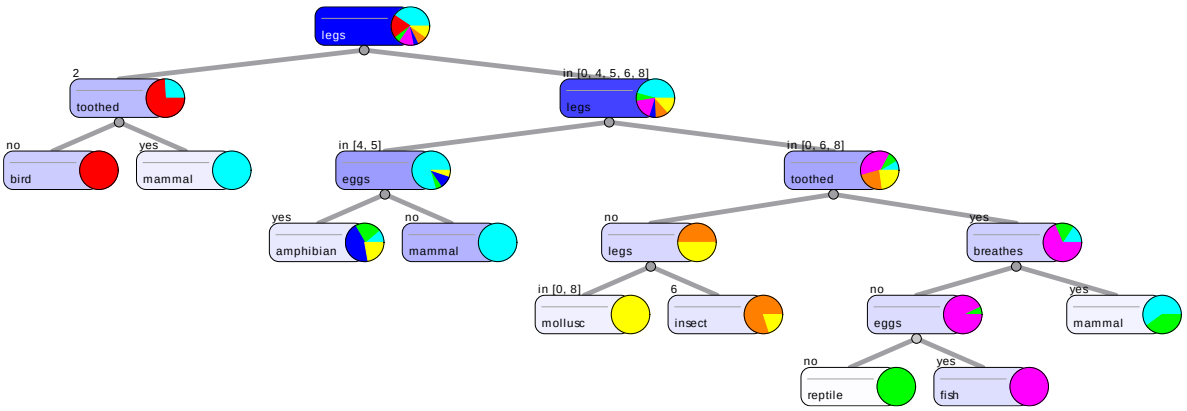


Figure 17: A tree with 9 leaves and branching factor 2 (same as the tree in Figure 5) learned using a subset of attributes with less evident relation with the class.

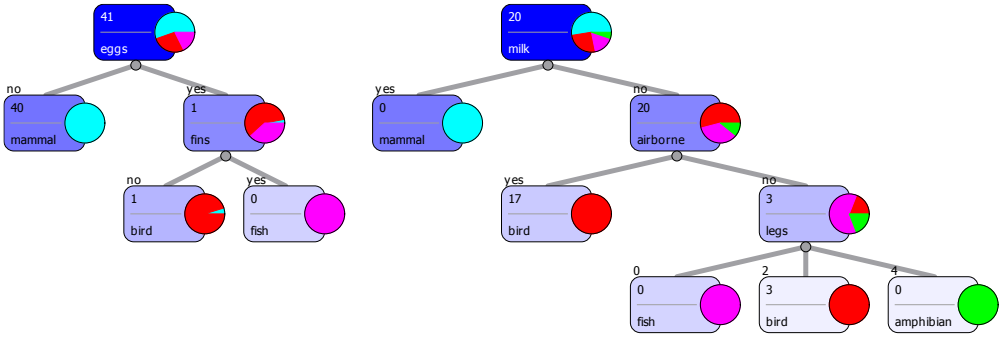


Figure 18: Small trees showing an unusual property of mammals (left) and birds (right).

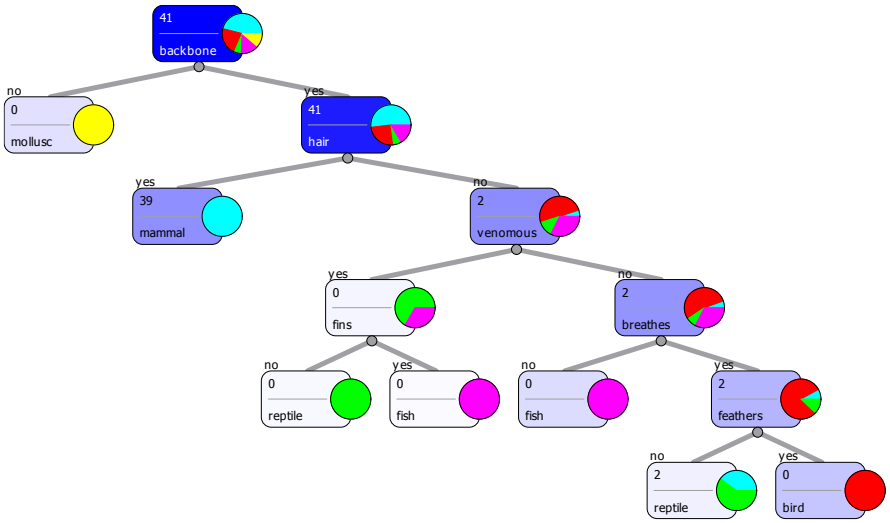


Figure 19: Medium tree showing unusual properties of mammals and fish (numbers in nodes correspond to the number of mammals in each node).

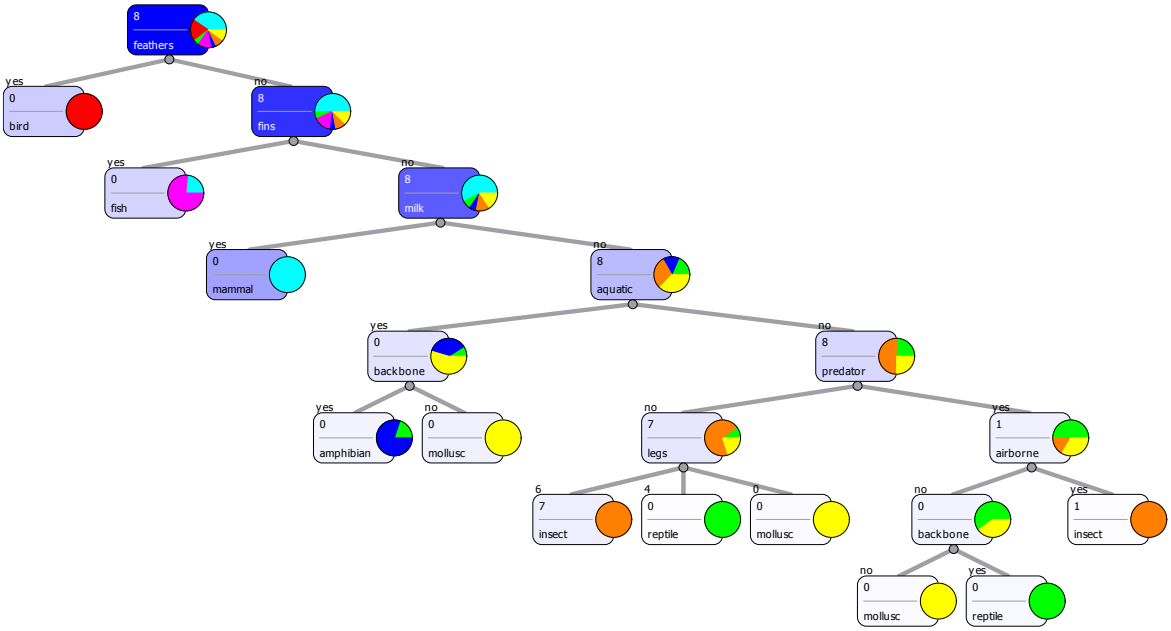


Figure 20: Big tree showing unusual properties of mammals, insects, and reptiles (numbers in nodes correspond to the number of insects in each node).

Assessing the potentials of cloud-native applications

Uroš Mesojedec[1,2] , Zoran Levnajić[1]

(1) Faculty of Information Studies in Novo mesto, Novo mesto, Slovenia

(2) T-media LLC, Novo mesto, Slovenia

Abstract: *Cloud computing represents the future of information technology as an exciting new shared platform. As with any new platform, success of the cloud computing is dependent on the availability of the adequate applications. While there are benefits of using traditional applications in the cloud, its real potential lies in specialised applications, developed and used with cloud in mind from the start. We call them "cloud-native" applications, and devote this paper to analysing the state-of-the-art in their development. We also present the benefits of cloud for novel approaches to application development. With the proper use of the proven architectural patterns, new breed of loosely coupled cloud applications with accessible user interfaces and enormous back-end processing power can be enabled. This can benefit even existing powerful software tools like GraphCrunch2, whose case we study in more detail.*

Keywords*: cloud computing, software application, patterns, network analysis*

1 Introduction

Cloud computing (or "the cloud" for short) is internet-based computing characterized by networking remote resources in order to allow for optimal computing accessibility and centralized data storage [1]. This enables the users to have online need-based allocation of large-scale computer services and computing resources. The concept of sharing the resources to achieve coherence and economies of scale is similar to the concept of utility (such as power grid). For example, a cloud facility that serves European users during European business hours with a specific application, may reallocate the same resources to serve North American users during North America's business hours with a different application, thus optimizing the resources. The term "cloud" is a metaphor of resources being available but hidden from the user 'in the cloud' (see Fig 1).

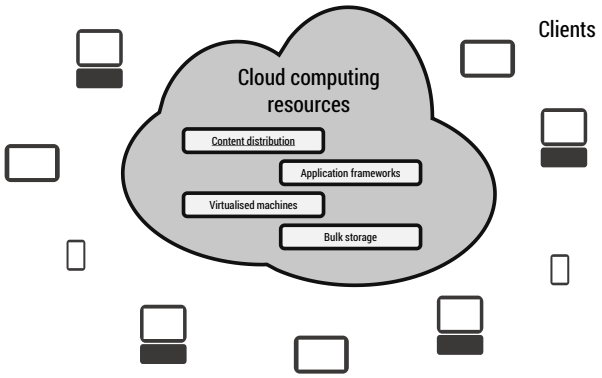


Figure Error! No sequence specified. *The metaphor of resources being hidden 'in the cloud'.*

Given our increasing everyday dependence on IT services, this novel evolution of computer technology might have an enormous and global impact [2]. Like the Web has transformed means of communication and doing business, its technological underpinnings are beginning to emerge as an alternative to in-house IT systems. Not just the technology progress but also new business models play a key role in propelling this new way of using IT infrastructure. Cloud enables big scale, parallel processing of very large data sets for everyone with access to Internet. It in general requires no initial investments in expensive equipment, but rather relies on the established and proven cloud infrastructure, available for (increasingly cheap) rent. Non-cloud systems are responsibility of the organisation itself, which is rarely adequately equipped for complex IT tasks. Their own systems are therefore complicated to design and maintain. Moreover, they usually have many other disadvantages such as not scaling well, which encourages organizations to switch to cloud as much as possible [3].

On the front of the personal users, cloud approach increasingly gains ground, as photos and music is easier to keep online than offline. In fact, even a superficial look at the current popular application landscape reveals a range of cloud-first applications that are enormously popular [4]. One such example is Dropbox [5], file synchronisation utility, which

allows to keep files online (for free up to certain size) - its core purpose is actually enabled by cloud. It is a cleanly separated client/server application, with clients for many popular platforms and cloud back-end that enables quick and easy synchronised storage. It is worth mentioning that Dropbox is running on Amazon's cloud infrastructure, meaning even the company itself took advantage of the cloud business model and avoided purchasing its own cloud.

The basic technical idea of cloud is leasing the resources which are provided as a general utility [6]. Cloud also redefines the roles of service provider into more specific infrastructure providers at the base and service providers on top. Several compelling features make cloud attractive: no up-front investment, lower operating costs, high scalability, easy access, reduction of business risks and maintenance costs. As with any new platform, overall success of the cloud is dependent on the availability of the right applications. There are benefits and trade-offs of using traditional applications in the cloud, but the real potential appears to be the employment of specialised applications and software, developed and adjusted particularly for cloud usage from the beginning.

The key focus of this paper is the analysis of current stage of the development of exactly those kinds of applications, which we term "cloud-native" applications, since they were never designed nor intended for usage outside cloud. We will present and discuss the benefits of the cloud environment and its efficient use for a novel approach to application development, done by using proven software development patterns combined with cloud advantages. Finally, we will specifically discuss the example of currently non-cloud application called GraphCrunch (a tool for scientific network analysis of biological data) [7] and pinpoint the advantages of its "cloudification" and ultimate conversion into a cloud-native application.

2 Cloud Environment

Services supporting our daily lives and improving its comfort, evolve and have always evolved in parallel with our societies. We can hardly imagine modern life without the utility services such as running water, electricity, telecommunications or heating. With improvement of standard of living, particularly in the Western countries, this list has been growing over the past decades, to eventually include Internet. With the expanding presence of IT in essentially all aspects of human activity, its absence becomes as unthinkable as the absence of any other utility from the list. There is in fact a growing need to organise the IT services to be exactly as yet another *utility*, simply available everywhere to everyone [8]. Recent developments in IT are converging to this goal, with cloud as the main proponent [9, 10].

Let us draw some parallels to a well-known utility. We all rely on a constant and high quality supply of electric energy. Immediately upon their invention, electric power plants used direct current, which forced their positioning close to the consumers. The invention of alternating current enabled long distance transfers of electric power with the transformer as key technology enabling it. This eventually allowed modern usage of electric power, whose key advantage is robustness to distance and involvement of the end user (consumer). History of electric power usage is largely analogous to what is expected to become the history of computing services, of which the Internet is a prime example. As transformers enable voltage conversions and utilization of the long distance transfer of energy, so are the routers enabling connection of local networks to the Internet. Eventually, the cloud is expected to allow for the same kind of robustness to the Internet usage, similar to that reached by the electric power grids.

Cloud can be seen as an easy to use computing resource that we either rent or build. Ubiquitous network infrastructure enables users to easily access and utilise this resources without the need for deeper understanding of cloud inner workings and its maintenance. Applications can therefore be easier to use and with more processing power. In particular, the web interface was proven to be very approachable and easy to use in a range of practical scenarios. On the other hand, regardless of their expertise in certain fields (science, medicine etc.), some users still lack the skills needed to efficiently utilize a specific powerful software package. Cloud has the potential to alleviate this and enable a wide and easy use of powerful software for everyone. This can in turn allow for a wider range of people to use Internet for their business and personal needs, regardless of their age, education level or knowledge of (foreign) languages.

Cloud computing systems transfer the responsibility of designing, running and maintaining arbitrarily complex IT systems to the outside party and allows regular users to easily use its applications, most commonly through a simple web browser. User has no knowledge and no worries about the issues related to maintenance of these systems, which enables him/her to enjoy the benefits of large computing power in the background cloud, accessed via simplified interfaces. On the back-end, cloud systems maintenance benefits from the economies of scale. Successful cloud providers are able to offer better service for less cost by providing specialised service for multitude of customers. They are in position to optimise all resources, from the electric power to the computing itself. Also, these systems are built to be fault tolerant and easy to maintain with automated recovery processes. Highly qualified maintenance teams are able to quickly solve potentially very complicated disruptions of services. It is very expensive for a non-cloud organisation to have such an expertise available in-house, which even if affordable, would be mostly underutilised.

There are many technical definitions of what cloud computing is or consists of. In [11] we find a layered model of cloud computing architecture (see Fig 2). It is structured into hardware, infrastructure, platforms and application layer, with different service models also being visible. At the lower end, both hardware and infrastructure layer, which include bare

metal resources and also their virtualization, can be exposed as an Infrastructure as a Service (IaaS). Above that, platform layer, including development frameworks, can be exposed as a Platform as a Service (PaaS). Even higher, application layer with finalized, user friendly interfaces, can be exposed as a Software as a Service (SaaS). Such architecture is modular and loosely coupled, which enables each layer to evolve on its own. All these different layers of exposure shield the user from the technicalities and finer details of the providers' technical solutions.

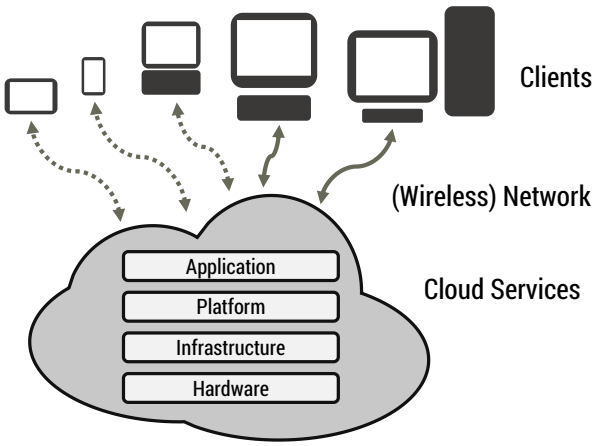


Figure 2 *Illustration of the cloud system layers defining its architecture.*

Using cloud services at the top layer of cloud architecture (SaaS) usually requires only a web browser. After years of competition among the browsers ('browser wars'), the prevalence of widely used software services enabled informal common functionality specification, which majority of web browser providers are required to supply without hindrance of proprietary enhancements and browser-specific bugs. Usage patterns strongly preferred the open web [12].

What we observed at the SaaS level is also a necessity for all other exposures, if they want to reach the utility status. Looking at the PaaS level, there are cloud resources, exposed at a platform which can be programmed into automated solutions. If we want to achieve utility status of cloud even at this level, some common programming frameworks should be available and as such would represent independent clients which would enable largest audience to participate on every cloud platform available. Embracing open standards, especially in cloud computing, which itself was enabled with open Internet protocols, is better long-term strategy, not only for users, but also for providers. Innovation happens above the common standardised platform and is available to a largest audience possible.

Software development is also affected by the cloud proliferation. There are different models to utilise the cloud for software applications. It is possible to host certain types of "traditional" applications (mostly) unmodified. In this scenario we only gain infrastructural benefits of a cloud system. On the other hand, cloud-native applications can gain all of the benefits of such systems. Cloud-native application should be written for the platform layer of cloud service and should be aware and actively use cloud specific services, such as high-speed connectivity, abundant processing resources and storage capacity on demand. This features are commonly not available for non-cloud applications, making cloud-native applications inherently novel and different.

3 Cloud Applications and the Process of Cloudification

Standard architectural patterns are commonly used in the development of any new software. As these development projects are usually complex, following the proven scheme of patterns can relieve some of the inherent risks of such endeavours and possibly facilitate the task of developers. Any developer wants to efficiently manage the complexity of all of the software parts, including those that we rely on every day.

Model View Controller (MVC) [13, 14] is one of the most successful architectural patterns, with a huge impact on software engineering since its introduction in the late 1970's. Its main purpose to establish a clear separation between the building blocks of the completed software project. The peculiarity of MVC pattern scheme, is that it fits nicely into the cloud paradigm, because it clearly separates the building block that faces the user (called 'View') from the data structures in the back-end (termed 'Model'). It also integrates two of them into a cohesive unity by connecting them through another block (called 'Controller', see Fig 3). These three blocks have their natural counterparts in the cloud system. Web interface corresponds to the View, Model is analogous to the data storage, while the Controller is the software 'glue' running in the platform layer of a given cloud environment. A variety of software projects that pre-date cloud successfully relied on MVC pattern. Yet its real benefits can perhaps be best exploited when employed to develop the cloud applications, in particular cloud-native applications. We could say that MVC is one of the better approaches in cloudification of existing applications. There are existing applications that utilize the MVC pattern in other paradigms,

e.g. desktop software with graphical user interface. For such applications the task of cloudification is made easier since they already employ the separation in application logic (by the MVC pattern) that is necessary for successful porting into the cloud paradigm.

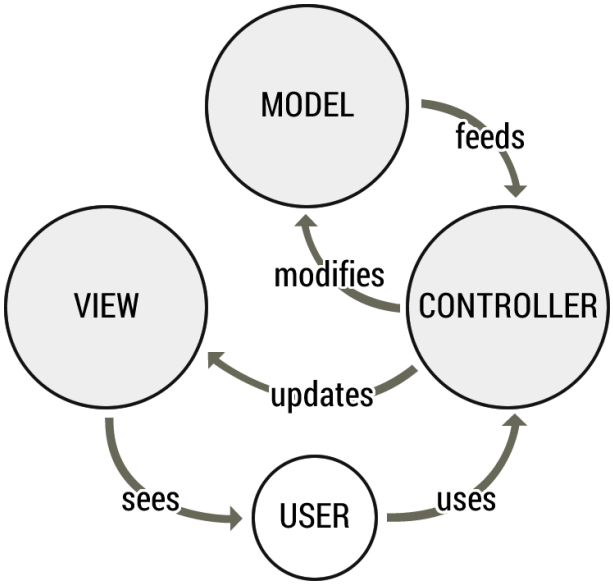


Figure 3 *Schematics of elements and interactions in the MVC architectural pattern.*

Clear separation between the data, processing logic, and the presentation is a natural state of any cloud application. Data is available from different sources, usually accessible through the resource location, for instance uniform resource locator (URL) address. This represents the Model of our pattern. Processing the data in our pattern is done by the Controller. Controllers in the cloud environment are commonly running in virtualised computing resources in a datacentre. To manage the application environment and present the results of a Controller we need the View. Progress of the web technologies enabled almost any View to be implemented as a web application. Therefore we can use arbitrary complex cloud application just through the web interface which is available on almost any networked device, from the desktop personal computers to the mobile smartphones. Employing the MVC pattern is therefore almost natural way of building cloud-native applications, which are loosely coupled, client-independent [16], very powerful, and at the same time open to further collaboration [17].

We take a closer look to an example of a successful cloud application. One of the most commonly used pieces of software is a simple spreadsheet, interactive cross-calculated table, allowing to easily write and use a table with numbers or words. Historical examples of a desktop spreadsheets are VisiCalc, Lotus 1-2-3, and in the recent years most commonly known Microsoft Excel [18]. Upgrading such application in the cloud was no easy task, but current solutions not only caught up with the capabilities of for example desktop versions, but have surpassed them on many levels.

There are many widely known benefits of cloud approach, e.g. user can access its data form many clients, collaboration is easier etc. Direct benefits of employing the MVC pattern are also evident. Efficient large scale data processing is possible as the Model can by any (big) data repository somewhere else in the cloud and the Controller can utilize many (virtualized) processor cores in the cloud data centre. Cost of infrastructure needed for large scale processing is minimised through cloud model, as we don't have to invest up front in the equipment capable of handling peak loads but we only rent such capabilities when needed. Also, simplified user interface through web based Views enables even non-experts in the field to efficiently use most of the application capabilities. Therefore, clearly separated cloud spreadsheet [19] by the MVC pattern has a central data model which is the same for all of the clients and users. This means collaboration is part of the pattern on which application was build. It can support simultaneous edits of many users, who can comment and interactively cooperate on the data. The entire history of edits is stored, allowing to revert to any past version, should any date become lost in the process. It becomes straightforward to use any URL-accessible data and import it into our calculations or use it while it updates in real time. Example of such multi-user spreadsheet can be readily accessed for free via Google Drive, for instance. Cloud capabilities are already in this respect an order of magnitude greater than that of any desktop class system. Finally, progress in web technologies made possible to develop very user-friendly View options, thus making the View practically equivalent to the simple web browser.

4 Large Network Analysis

We live in the age of Big Data, where increasing accumulation of new data call for new approaches in analysing them. Networks (graphs) are a powerful mathematical framework that allows to elegantly represent and study complex systems and in general complex datasets. Complex systems on the other hand, can be found on all scales in nature and in society, ranging from information, social, physical, transportation, and biological systems. Among many different types of biological networks, protein-protein interaction (PPI) are possibly the most interesting, where proteins within cell are represented as nodes, while their chemical interactions are modelled as links. The network analysis software GraphCrunch 2 (GC2) is a tool that implements the latest research on biological network analysis, in particular PPI networks [20]. It is a substantial upgrade of original GraphCrunch which implements the most used random network models and compares them to the data networks, analysing specific network properties. GC2 also implements GRAph ALigner (GRAAL) algorithm for topological network alignment which can expose large, contiguous, dense regions of topological and functional similarities. In this regard GRAAL surpasses any other existing tool. GC2 is already a very efficient and useful tool, even for experts without extensive knowledge of networks and network modelling (see Table 1).

Software package	*Graph properties*	*# of models*	*Graphlets*	*Visualization*	*Clustering*	*GNA*
GraphCrunch 2	Yes	7	Yes	Yes (Results)	Yes	Yes
GraphCrunch	Yes	5	Yes	Yes (Results)	No	No
Cytoscape	Yes	6	Limited	Yes	Yes	No
Visant	Yes	1	No	Yes	No	No
mFinder	No	3	No	Yes (mDraw)	No	No
MAVisto	No	1	No	Yes	No	No
FANMOD	No	3	No	Yes	No	No
tYNA	Yes	0	No	Yes	No	No
pajek	Yes	2	No	Yes	Yes	No
IsoRank	No	0	No	No	No	Yes
Graemlin	No	0	No	No	No	Yes
GraphM	No	0	No	No	No	Yes

***Table** 1 Comparison of software tools for biological network analysis (table source [20])*

However, GC2 could benefit even more from the cloudification. The challenges for further advancement of GC2 capabilities and its wider use can be efficiently addressed by porting the application into the cloud environment. By using the above mentioned MVC pattern separation of application modules large community of network experts, users from other science domains and also software developers can be engaged on different levels. Also, cloud environment will enable easier use of large data models, available through URL. Cloud datacentre capabilities can drastically expand software tool computing power without the need for up-front investments in hardware. If large processing power is required for especially complex network analysis, it can simply be rented for a reasonable price, driven lower every day through harsh competitive ecosystem of different cloud providers [21]. Clear separation of application modules through the MVC model could also expand GC2 popularity with developer community. For example, as software is open sourced, experts in the web interface domain could be engaged to provide just View plug-ins, without even handling internal Controller logic with raw network analysis capabilities. The same can be said for network experts, which can provide Controller expansion, without the need of interacting with the View parts.

We should not forget of the social component of cloud available software. Collaboration is made much easier. For example results can be shared simply by emailing the URL of the specific View. This opens great opportunities of further analysis, discussion and expansion of results by using all of the popular social tools. Cloud also encourages real-time collaboration [22]. Many experts can work on the same Model simultaneously either through different Views or even using the same View if their research is tied.

5 Disussion and Conclusions

We have presented the benefits of a cloud environment for a novel approach to application development. Cloud has a large potential to revolutionize not only IT architectures but also software development patterns and application usage.

Cloud has the right potential to enable global computing infrastructure as another utility service. But it is reliant on other infrastructure. It is obvious it needs electric power, which although not globally available, is a common utility service. Also mobile, specifically battery technology, can alleviate some of the issues of using cloud applications in areas with not so reliable power sources. Another crucial infrastructure technology for cloud is fast and reliable

telecommunication network. This is an area of large interest of cloud businesses [23], because majority of human population still cannot enjoy broadband connectivity. We can reasonably expect this situation to improve drastically in the following years.

With the proper use of the proven architectural patterns, such as the MVC pattern, we can enable new breed of loosely coupled cloud applications with accessible user interfaces and enormous back end processing power which is available from the cloud datacentres. This opens the path for solving the entirely new class of problems and also enables easier collaboration and knowledge sharing.

This can be demonstrated on the specific example of porting existing powerful application, like GraphCrunch 2, into the cloud paradigm. Even without functional expansion, there are many benefits of porting it to the cloud. With the port available, there are many further possibilities of expanding application functionality and also larger and easier use, collaboration and sharing of the research results.

6 References

[1] Wikipedia: Cloud Computing. Retrieved September 2014 from http://en.wikipedia.org/wiki/Cloud_computing

[2] Aljabre, A., Cloud computing for increased business value. *International Journal of Business and Social Science*, 3(1), 2012, pp. 234-239.

[3] Youseff, L., Butrico, M., Da Silva, D., Toward a unified ontology of cloud computing. *Grid Computing Environments Workshop*, 2008. GCE'08 IEEEE, pp. 1-10

[4] McKendrick, J., 20 Most Popular Cloud-Based Apps Downloaded into Enterprises, Forbes.com, April 2013. Retrieved September 2014 from http://www.forbes.com/sites/joemckendrick/2013/03/27/20-most-popular-cloud-based-apps-downloaded-into-enterprises/

[5] Drago, I., et al., Inside dropbox: understanding personal cloud storage services, *Proceedings of the 2012 ACM conference on Internet measurement conference*, 2012, pp. 481-494.

[6] Mell, P., Grance, T., Cloud Computing Synopsis and Recommendations. *NIST Special Publication 800-146*, National Institute of Standards and Technology, U.S. Department of Commerce, May 2011. Retrieved September 2014 from http://csrc.nist.gov/publications/drafts/800-146/Draft-NIST-SP800-146.pdf

[7] Milenković, T., Lai, J., Pržulj, N., GraphCrunch: a tool for large network analyses, *BMC bioinformatics*, 9(1), 2008.

[8] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., Brandic, I., Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, 25(6), 2009, pp. 599-616.

[9] Foster, I., Zhao, Y., Raicu, I., Lu, S., Cloud computing and grid computing 360-degree compared, *Grid Computing Environments Workshop*, GCE'08, 2008, pp. 1-10.

[10] Rappa, M. A., The utility business model and the future of computing services, *IBM Systems Journal*, 43(1), 2004, pp. 32-42.

[11] Zhang, Q., Cheng, L., Boutaba, R., Cloud computing: state-of-the-art and research challenges, *Journal of Internet Services and Applications*, 1(1), 2010, pp. 7-18.

[12] West, J., Mace, M., Browsing as the killer app: Explaining the rapid success of Apple's iPhone, *Telecommunications Policy*, 34(5), 2010, pp. 270-286.

[13] Reenskaug, T., The Original MVC Reports. *Xerox Palo Alto Research Laboratory*, PARC, 1978.

[14] Krasner, G.E., Pope, S.T., A description of the model-view-controller user interface paradigm in the smalltalk-80 system, *Journal of object oriented programming* 1.3, 1988, pp. 26-49.

[15] Burbeck, S., Applications Programming in Smalltalk-80: How to Use Model-View-Controller (MVC). Softsmarts, Inc., 1987. Retrieved September 2014 from http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html

[16] Mesojedec, Uroš. Client-Independence as the Critical Step Towards a New Level of Cloud Computing, *Proceedings of the 7th European Computing Conference* (ECC'13), Dubrovnik, Croatia, June 25-27, 2013, (Recent Advances in Computer Engineering Series, 13). WSEAS Press, cop. 2013, pp. 179-184.

[17] Manjunatha, A., Ranabahu, A., Sheth, A., Thirunarayan, K., Power of Clouds In Your Pocket: An Efficient Approach for Cloud Mobile Hybrid Application Development. *2nd IEEE International Conference on Cloud Computing Technology and Science*, IEEE, 2010, pp. 496–503.

[18] Wikipedia: Spreadsheet. Retrieved September 2014 from http://en.wikipedia.org/wiki/Spreadsheet

[19] Herrick, D. R. Google this!: using Google apps for collaboration and productivity, *Proceedings of the 37th annual ACM SIGUCCS fall conference*, 2009, pp. 55-64.

[20] Kuchaiev, O., Stevanović, A., Hayes, W., Pržulj, N. GraphCrunch 2: software tool for network modeling, alignment and clustering. *BMC bioinformatics*, 12(1), 24, 2011.

[21] Ograph, B. T., and Morgens, Y.R., Cloud computing. *Communications of the ACM 51.7*, 2008.

[22] Graham, M., Cloud collaboration: Peer-production and the engineering of the internet. *Engineering earth*. Springer Netherlands, 2011, pp. 67-83.

[23] Google: Project Loon. Retrieved September 2014 from http://www.google.com/loon/

Alternative way of evaluation of air pollution levels

Biljana Mileva Boshkoska
Faculty of Information Studies
Sevno 13, 8000 Novo mesto, Slovenia
Biljana.mileva@fis.unm.si

Abstract: *Air pollution is a constant environmental problem, which introduces significant costs primarily to the health of the society, as well as to the ecosystem, and the economy. The main difficulty in the modern modelling approaches is their interpretation by the wider public. In this paper, we provide a preliminary modeling approach that can be used for aggregation of the different air quality pollutants into one air quality measure described qualitatively and hence can be understood by the wider population. We used the modeling approach DEX that allows both scientific notation of the model and easy-to-interoperate evaluation results by the wider public.*

Key Words*: air pollution evaluation, air quality, DEX modelling*

1 Introduction

The media reports may lead us that the problem of ambient air pollution arose in the second half of the last century. However, the pollution of ambient air is not a new phenomena in the history of mankind. Obviously even for the caveman the fire lightning had its consequences (1). There are historical data on destruction of plants as a result of the furnaces since the Roman Empire. However, the type of pollution of ambient air to which people have been exposed have changed throughout history, but the problem has been known for a long time, and cought the interest of the public especially in the 14th century when people first began to use coal for heating in their homes (2).

Today, the problems due to air pollution can be found in several areas. Air pollution is a constant environmental problem, which introduces significant costs to the health of the society, the ecosystem (16), and the economy.

The most important and obvious problems to air pollution are the direct effects on human health. The three pollutants that are recognized to most significantly affect human health are Particulate matter, nitrogen dioxide and ground-level ozone (17, 18). Recent research indicates that small particles (PM2.5) in the air caused about 450,000 premature deaths within the 27 EU countries in the year 2005. Another 20,000 premature deaths was caused by ground-level ozone (19). Small particles were also responsible for around 100,000 serious hospital admissions in the EU25, and a much larger number of less serious effects, for example some 30 million respiratory medication use days and several hundred million restricted activity days.

Next in line are the effects and damage to our environment such as the acidification of

lakes, including the soil deforestation, eutrophication, ozone at ground level or crop damage.

Finally, the problems of air pollution overlap with other complex environmental issues as such congestion and mobility, landuse and global warming.

Today, science is concerned with modeling of parameters of the ambient air in order to investigate or to protect and improve the environment in which we live (3, 4, 5). Many researches deal with modeling the air quality in order to fill in the gaps of missing data, or to predict the air quality (6, 7, 8, 9, 10, 11, 12, 13, 14, 15). These models are usually based on historical data, they use complex algorithms and are usually understood by a small population of the relevant domain researchers. In this paper we aim at providing a tool for determining the air pollution level as an aggregated value. Moreover, the tool can be easily used by a wide population, without the need for an expert level understanding of the air pollution processes or modeling techniques. In particular, this modeling approach can be used for real-time data on air pollution levels. The background motivation of this work arises from the fact that web cites usually provide information regarding different pollutant levels, without an aggregated information about the overall air pollution. For example, if we check the EEA web cite http://www.eea.europa.eu/themes/air/air-quality/map/real-time-map, we get maps for four unvalidated real-time air quality pollutants across Europe: ozone (O3), nitrogen dioxide (NO2), Particulate Matter 10 (PM10) and Sulfur dioxide (SO2). However, there is no aggregated map of the overall air quality.

In this paper we provide an preliminary modeling approach that can be used for aggregation of the different air quality pollutants into one air quality measure described qualitatively and hence can be understood by the wider population.

2 Methodology

In order to protect the human health and vegetation, the Council directive 1999/30/EC has been adopted which defines the limit values of air pollutants and the dates until which these had to be met. Despite these measures, the air quality in European countries frequently breaches the limits of aloud concentrations of ambient air parameters (20, 21). Reasons for increased air pollution may be found in the natural weather conditions as well as in the economic crises (22).

The limit values of the parameters of ambient air are given in Table 1 - Table 4.

Table 1 Limit value and the alert threshold for sulphur dioxide (23)

	Averaging period	Limit value micrograms/m3	Comment
Hourly limit value for the protection of human health	1 hour	350	Not to be exceeded more than 24 times a calendar year
Daily limit value for the protection of human health	24 hours	125	Not to be exceeded more than 3 times a calendar year
Limit value for the protection of	Calendar year and winter (1.9 – 31.3)	20	

ecosystems			

Table 2 Limit value and the alert threshold for NO2 and oxides of nitrogen (23)

	Averaging period	Limit value micrograms/m3	Comment
Hourly limit value for the protection of human health	1 hour	200 NOx	Not to be exceeded more than 24 times a calendar year
Daily limit value for the protection of human health	24 hours	40 NO2	Not to be exceeded more than 3 times a calendar year
Limit value for the protection of vegetation	Calendar year	30 NOx	

Table 3 Limit value and the alert threshold for particulate matter (23)

	Averaging period	Limit value micrograms/m3	Comment
24-hour limit value for the protection of human health	24 hours	50	Not to be exceeded more than 7 times a calendar year
Anual limit value for the protection of human health	Calendar year	20	

Table 4 Target value for ozone (23)

	Averaging period	Target value micrograms/m3	Comment
Protection of human health	Maximum daily eight-hour mean	120	Not to be exceeded more than 25 days per calendar year averaged over three years
Protection of vegetation	May to July	AOT40 (calculated from 1 h values)	

The values of the air pollutants are continues measurements which can be converted into qualitative ones by using the guidelines in (23). These conversion levels are given in Table 5 (http://www.eea.europa.eu/themes/air/air-quality/map/real-time-map):

Table 5 Quantitative to qualitative mapping of pollutants levels according to guidelines in (23)

Level of O3 in micrograms/ m3	Level of NOx (NO and NO2) in micrograms/m3	Level of PM (PM10 and PM2.5) in micrograms/m3	Level of SO2 in micrograms/m3	Qualitative description

0 - 60	0 - 50	0 - 20	0 – 50	Low
60 - 120	51 - 100	21 - 35	51 – 100	Slight
120 - 180	101 - 200	36 - 50	100 – 300	Moderate
180 - 240	201 - 400	51 - 65	301 – 500	High
Above 240	Above 400	Above 65	Above 500	Very high

2.1 Qualitative modeling with DEX

DEX (25) belongs to the group of qualitative multi-criteria decision making (MCDM) methods. In DEX, the qualitative attributes build a hierarchical structure which represents a decomposition of the decision problem into smaller, less complex and possibly easier to solve sub-problems. There are two types of attributes in DEX: basic attributes and aggregated ones. The former are the directly measurable attributes, also called input attributes, that are used for describing the options. The latter are obtained by aggregating the basic and/or other aggregated attributes. They represent the evaluations of the options. The hierarchical structure in DEX represents a tree. In the tree, attributes are structured so that there is only one path from each aggregate attribute to the root of the tree. The path contains the dependencies among attributes such that the higher-level attributes depend on their immediate descendants in the tree. This dependency is defined by a utility function. The higher-level attribute, its immediate descendants and the utility function form a qualitative decision table.

In DEX, the aggregation of the qualitative attributes into a qualitative class in each row in the decision table is interpreted as if-then rule. Specifically, the decision maker's preferences over the available options are given with the attribute that is called a qualitative class. Options that are almost equally preferred belong to the same qualitative class.

3 Results

The developed DEX model tree that is used for the evaluation of air quality is presented in *Figure 1*. The basic attributes in *Figure 1*are given with rectangles with curved edges, such as PM2.5, PM10, NO2, NO and O3. The aggregated ones are given with rectangles with sharp edges, such as PM, NOx, O3 and Air Quailty. The value scales of each attribute are given in Table 5, which is obtained from the implementation of the DEX model for air quality assessment in the computer program DEXi (24).

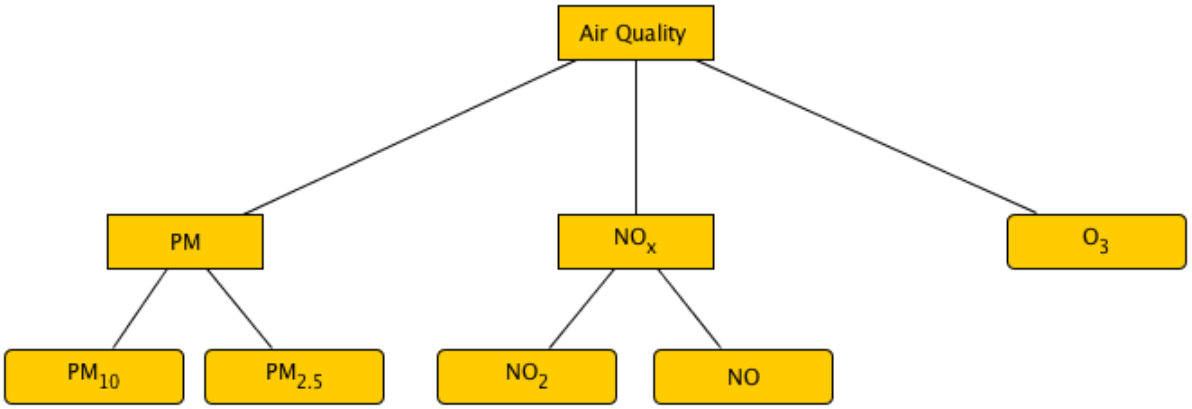


Figure 1 DEXi model tree for evaluation of air pollution

The evaluation of the air quality is performed by using aggregation functions given in a tabular format. For example, the aggregation of the qualitative attributes PM10 and PM2.5 into a qualitative attribute PM is given with Table 6. Such tables are created for all aggregated attributes presented in Figure 1. These aggregation functions can also be read as *if-then* rules. For example, the first row in Table 6 can be read as: *if PM10 is Low AND PM2.5 is Low then PM is Low.* This enables both practitioners and wider population to make use of the results. Practitioners obtain a scientific representation of the data, while the wider public obtains an easy to understand explanation about the air pollution.

Table 6 Aggregation of basic attributes PM10 and PM2.5 in qualitative attribute PM

PM10	PM2.5	PM
Low	Low	Low
Slight	Low	Slight
Moderate	Low	Moderate
High	Low	High
Very High	Low	Very High
Low	Slight	Slight
Slight	Slight	Slight
Moderate	Slight	Moderate
High	Slight	High
Very High	Slight	Very High
Low	Moderate	Moderate
Slight	Moderate	Moderate
Moderate	Moderate	Moderate
High	Moderate	High
Very High	Moderate	Very High

Low	High	High
Slight	High	High
Moderate	High	High
High	High	High
Very High	High	Very High
Low	Very High	Very High
Slight	Very High	Very High
Moderate	Very High	Very High
High	Very High	Very High
Very High	Very High	Very High

4 Conclusions

Air pollution introduces significant costs to the health of people, the ecosystem, and the economy. The main difficulty in the current modelling approaches is their interpretation by the wider public. In this paper, we provide a preliminary modelling approach that can be used for aggregation of the different air quality pollutants into one air quality measure described qualitatively. For modelling of the air pollution, the DEX methodology was used that allows both scientific notation of the model and easy-to-interoperate evaluation results by the wider public. The future work should include more parameters in the DEX model.

5 Acknowledgements

This work is supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund.

6 References

[1] Elvingson, P; Agren. C. Air and the environment. Elanders Infologistics AB. Molnlycke, Goteborg, Sweeden, 2004.
[2] Schnelle, K.B; Brown, C.A. Air Pollution Control – Handbook. CRC Press, 1997.
[3] Cheremisinoff, N.P. Handbook of Air Pollution Prevention and Control. Butterworth Heinemann, 2006.
[4] Boubel, R; Fox, D.L; Turner, B.D; Stern, A.C. Fundamentals of Air Pollution. Academic Press, 1994.
[5] Turner, B.D. Atmospheric dispersion estimates. Lewis Publishers, 1994.
[6] Rakotomamonjy, A. Variable Selection using Support Vector Machines based Criteria. Neurocomputing, 1-2 (55), 2003.
[7] Berkowicz, R; Winther, M; Ketzel, M. Traffic pollution modelling and emission data. Environmental Modelling & Software, 21: 454-460, 2006.
[8] Calori, G; Clemente, M; De Maria, R; Finardi, S; Lollobrigida, F; Tinarelli, G. Air quality integrated modelling in Turin urban area. Environmental Modeling & Software, 21(4):467-476, 2006.
[9] El-Halwagi, M.M. Pollution Prevention through Process Integration. Academic Press, 1997.
[10] Farias, F; ApSimon, H; Relative contributions from traffic and aircraft NOx emissions to exposure in West London. Environmental Modelling & Software, 21(4):477-485, 2006.
[11] Cabrera, G; Martinez, F; Mateosa, M; Tavera, V. Study of the evolution of air

pollution in Salamanca (Spain) along a five-year period (1994–1998) using HJ-Biplot simultaneous representation analysis. Environmental Modelling & Software, 21(1):61-68, 2006.
[12] Raimondo, G; Montuori, A; Moniaci, W; Pasero, E; Almkvist, E. A machine learning tool to forcast PM10 level. Fifth Conference on Artificial Intelligence Applications to Environmental Science, San Antonio, Texas, USA, 2007.
[13] Canu, S; Rakotomamonjy, A. Ozone peak and pollution forecasting using support vectors. IFAC Workshop on Environmental Modelling, Yokohama, Japan, 2001.
[14] Neofytou, P; Venetsanos, A.G; Rafailidis, C; Bartzis, J.G. Numerical investigation of the pollution dispersion in an urban street canyon. Environmental Modelling & Software, 21(4):525-531, 2006.
[15] Sokhi, R. S; San Joseb, R; Kitwiroon, N; Fragkou, E; Perez, J.L; Middleton, D. R. Prediction of ozone levels in London using the MM5–CMAQ modelling system. Environmental Modelling & Software, 21(4):566-576, 2006.
[16] Effects of air pollution on European ecosystems, Past and future exposure of European freshwater and terrestrial habitats to acidifying and eutrophying air pollutants, EEA Technical report, No 11/2014 (http://www.eea.europa.eu/publications/effects-of-air-pollution-on) last checked: August 12th, 2014.
[17] European Environment Agency, Air pollution – European, http://www.eea.europa.eu/themes/air/intro, last checked: September 15th, 2014.
[18] Air Climate & Climate Secretariat, Toxic air in Europe, http://www.airclim.org/toxic-air-europe, last checked: September 15th, 2014.
[19] Air Climate & Climate Secretariat, Air Quality, http://www.airclim.org/air-quality, last checked: September 15th, 2014.
[20] From The Atlantic Citylab, Western Europe's Mild March http://www.citylab.com/commute/2014/03/western-europes-mild-march-has-led-air-quality-crisis-france/8640/, last checked: September 15th, 2014.
[21] UK Air pollution, http://blogs.egu.eu/hazeblog/2014/03/14/uk-air-pollution-march-2014/, last checked: September 15th, 2014.
[22] The Christian Science Monitor, How Greece's economic crisis filled the Athens sky with smog, http://www.csmonitor.com/World/Europe/2014/0212/How-Greece-s-economic-crisis-filled-the-Athens-sky-with-smog, last checked: September 15th, 2014.
[23] Directive 2008/50/ec of the European parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe, http://eur-lex.europa.eu/, last checked: September 15th, 2014.
[24] Bohanec, M. DEXi: Program for Multi-Attribute Decision Making: User's manual: version 4.00. IJS Report DP-11340, Jožef Stefan Institute, Ljubljana, 2013.
[25] Bohanec, M; Rajkovic, V. DEX: An expert system shell for decision support. Sistemica 1, 145–157, 1990.

Case Study: Web Clipping and Sentiment Analysis

Jože Bučar, Janez Povh
Faculty of Information Studies
University of Novo mesto
Sevno 13, 8000 Novo mesto, Slovenia
{joze.bucar, janez.povh}@fis.unm.si

Abstract: *Web clipping deals with retrieving and extracting text and graphic components from web pages and efficient display on a hand-held web appliance. The projects are conducted in collaboration with company Nevtron & Company, d.o.o. that manages the leading IT portal in Slovenia and produces news media, whose role involves experimental testing of proposed solutions. Our aim is to develop web application for data retrieval to find out where, when and in what form information appears on the web and to determine the response from readers. We describe an approach that automates fast and effective collection of data which is crucial for future editorial and business decisions. We have developed solutions that provide a better flexibility to users in selecting, retrieving, extracting and tracking information in publicly available publications on the web.*

Key Words*: web clipping, information retrieval, sentiment analysis, text mining, web mining*

1 Introduction

An enormous quantity of data is generated on the web daily. We are practically deluged by all kinds of data – scientific, medical, financial, historical, health care, demographic, business, and other. Usually, there are not enough human resources to examine this data. From this chaotic cluster of data we strive to obtain valuable information, which may significantly impact strategic decisions of both business and individuals in the future. The increasing interest in web content has attracted the collaboration among scientists from various fields such as computer science, data mining, machine learning, computational linguistics, graph theory, neural networks, sociology, and psychology. The ability to understand and gain knowledge from text is a key requirement in the goal of artificial intelligence researchers to create machines that simulate the most complex thinking machine in the universe: the human brain.

Relevant information about an organization, its structure, employees, activities, products and services can appear anywhere on the web. An increasing number of blogs, web sites, newsgroups, forums, chat rooms, etc. has allowed people to express and aggregate their feelings about products, services, events, popularity of political candidates more intensively [8]. Although information on the web can be either true or false, it has significant impact on public opinion and its response [11]. However, more and more business, sale, finance, and other companies are aware of people's opinion. In a constant battle for success businesses are looking to market their products, identify new opportunities and manage their reputations. The popularity of social media such as social networks, blogs and others has escalated interest in sentiment analysis [15].

Early text mining activities were concerned primarily with various forms of information retrieval and information summarization, such as indexes, abstracts, and grouping of documents. Later developments in the text mining focused on information extraction. The content and relationship information is extracted from a corpus of documents. These information extraction operations are performed by tagging each document for the presence of certain content facts or relationships between them [9]. Information extraction consists of an ordered series of steps designed to extract terms, attributes of the terms, facts,, and events [12]. Typical text mining tasks include: classification and categorization of texts (documents), topic detection, sentiment analysis, summarization (summary) of texts and the study of relationships between entities in the texts.

In most developed countries automated monitoring of information on the web and other media is an everyday routine, since it improves distinctness, ensures stability and reputation of organizations. There are several successful companies, offering comprehensive solutions from detecting, filtering, classifying, analysing and informing users such as Google Alerts [3], Web Clipping [14], etc... In Slovenia automated monitoring of information on the web is not yet widespread.

The projects Web Clipping and SEAN (Sentiment Analysis) are both developed in collaboration with prominent company Nevtron & Company d.o.o. (Ljubljana, Slovenia). The company is interested in an advanced web application for tracking and retrieving up to date information, and to set a level of importance of the phenomena for their (potential) clients. The purpose is to obtain up to date and accurate information about client's products and services (e.g. news, articles, etc.). In addition to where, when and in what way a certain piece of information appears, the company is also interested in the detecting importance of this information and user response. The company as well as its clients would like to know whether published news and articles were favourably accepted or not. The project is market-oriented and encourages the demand for services (organization and management with more information on the organization, product, service, etc.). Knowing what users think about organization, its products, and services has a huge potential because it improves company's future editor policy and its strategic business decisions. Proposed solution will enhance company's portfolio of services, and hopefully, the effort will contribute to penetration on international markets.

The rest of the paper is organized as follows: Section 2 introduces Architectural Framework, where we discuss about proposed approach and implemented solutions. Finally, the paper ends with Conclusions in Section 3.

2 Architectural Framework

The Figure 1 illustrates the architectural framework of our proposed solution, which is based on user studies and is divided into two components:

- Automatic web text retrieval and record to the database
- Sentiment analysis.

In the following sections we will present each component in greater detail. The web application involves a group of processes that require considerable memory and processing capabilities. Therefore, it will run on company's server and its performance is dependent on the capacities and limitations of that machine.

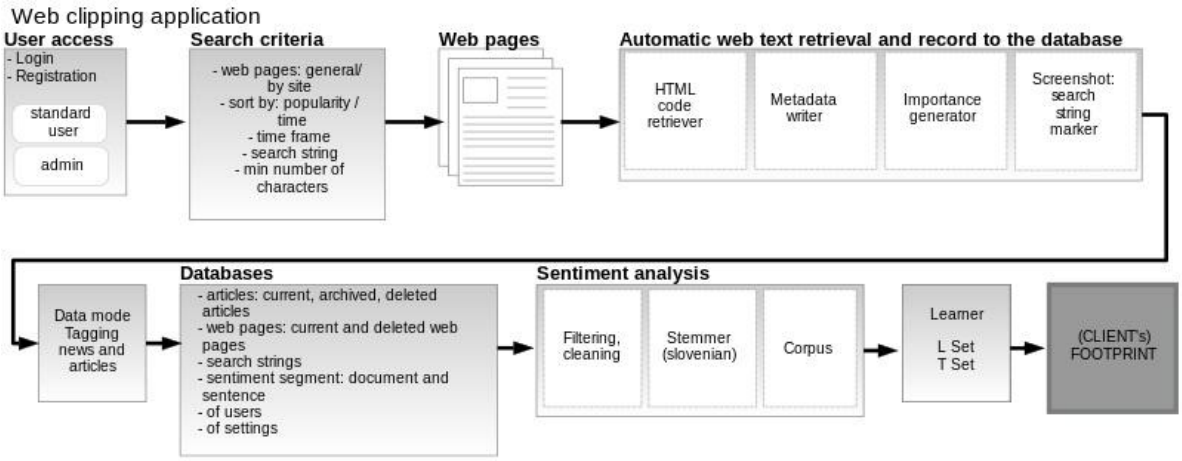


Figure 1: Architectural framework

2.1 Automatic web text retrieval and record to the database

This component is a part of Web Clipping project. Using constrains in search criteria this component allows users to specify what part of the contents of a web page is relevant to them. This component finds and obtains objects within web pages, enables editing and storing content and metadata in databases on company's server.

Initially we studied which of the existing search engines is suitable to deal with the challenge. We chose Google Custom Search Engine [4], which adjusts the search parameters to implement specified solutions. Within the project we have developed a module for automatic identification which contains:

- Login and registration form;
- Two-level access: administrator and standard user (with limited functionalities);
- Logout button and automatic logout after longer absence;
- The search feature;
- Traceability of posts (automatic process; search engine finds results according to a given input and search criteria);
- Built-in function to review and select one or multiple selection of results with an option to save each search result;
- Function for optimal HTML code recognition and retrieval;
- Function to retrieve (and edit) the entire HTML source code within web page, which is suitable in case of errors in HTML sources;
- Option for editing, storing, archiving, and deleting content or metadata;
- Advanced search within databases;
- Exporting results to pdf format, which includes predefined search criteria, important information about search results, web pages where content of search results occurs, screenshots with marked search string, and graphical display of results.

The search is carried out within search criteria:

- Search within Web pages or databases: User can select among default set of web pages, define a new set of web pages, or can search over all pages on the Web;
- Sort by popularity or time of publishing;

- Search within time frame: User can search without specifying a time frame, by entering the time frame manually via the key-board, or by selecting the time frame within a built-in calendar;
- Search string: User can enter a keyword, a set of keywords, or other specified string for the proposed search;
- Minimal number of characters: User can determine the length of a string in which the search string should appear.

HTML code retriever enables automatic recognition of HTML code. For this purpose we developed a universal text parser for all websites, and customized text parsers for the default set of web pages to enhance the acquisition of the content.

Metadata writer is bound to store information (into databases) about search results such as article ID, the date of entry into the data-base, the URL of the parent web page, the URL of news or article, its title, keywords, the name of a user who initiates search, importance of publication of the web page (TR and GL rank), and screenshot.

Importance generator generates the importance of the content based on traffic ranks. We used two ranks: Global rank (GR) and Local rank (LR) [1]. Global rank is an estimation of site's popularity, and local rank is an estimation of site's popularity in specific country. Both ranks are calculated using a combination of average daily site's visitors and page views from users over the past month. The site with the highest combination of visitors and page views is ranked the highest.

Marking the search string and capturing site's screenshot was accomplished by using PhantomJS. We were able to access and manipulate DOM objects within web pages. Since PhantomJS is using WebKit, a real layout and rendering engine, it can capture a web page as a screenshot. PhantomJS can render anything on the web page so it can be used to convert contents not only in HTML and CSS, but also in SVG and Canvas [10].

2.2 Sentiment analysis

This component is part of another research project SEAN and is still under construction. In this research we retrieved 198.187 textual documents from different Slovenian websites.

From the corpus of analysed unstructured data retrieved from the web, sentiment analysis will be performed:

- Filtering, cleaning and pre-processing of data;
- Manual annotation of randomly selected textual content will base on five-level Likert scale [7], where: 1 – very negative, 2 – negative, 3 – neutral, 4 – positive and 5 – very positive. To evaluate the process of annotation inter-rater and intra-rater reliability will be used;
- Tokenization, transcription, stop words, stemming, and lemmatization will be tried [5];
- Testing and integration of machine learning algorithms for the classification of sentiment; performance and evaluation of tested classification algorithms;
- Branding of the company (identification sentiment about the company, evaluating response to posts and analysis).

Classification of (web) texts is one of the key tasks in text mining. Automation of procedures for the purpose of classification of texts has thus become an important tool, which has contributed to more efficient work. Therefore, data miners use a variety of tools and a wide range of learning algorithms [13].

By labelling a sample of 640 documents (468 negative and 172 positive) we obtained a labelled corpus, which was used as a training set to train, test and evaluate classification techniques. Each document was represented as a bag of individual words (unigrams). All words containing capital letters were transformed into lowercase, also Terms frequency (TF), Term Frequency – Inverse Document Frequency (TF-IDF) were used. In the process of attribute selection we did not use stop words for Slovenian language, nor algorithms for stemming or lemmatization. We carried out a classification of web texts and performance of learning algorithms. In order to efficiently predict the category (positive or negative) of 640 documents we chose to test and evaluate two commonly used classification methods: Naïve Bayes [6] and Support Vector Machines [2]. As a result we found out that in case, when learning set consists of all documents, the Support Vector Machines correctly classified all documents, while the Naïve Bayes provided 84.38% accuracy. Estimated accuracies were practically over fitted to our data. In case, when we used 10-fold cross validation both methods produced similar results (Naïve Bayes with: 76.88% and Support Vector Machines with 75.63% accuracy).

3 Conclusions

In this paper we briefly introduced the results of two research projects: Web Clipping and SEAN. We proposed an approach and preliminary work within the web solution for selecting, retrieving, extracting and tracking information in publicly available publications on the web.

During development of web application we encountered some challenges, which we managed to overcome successfully. However, we had some issues with adaptation and implementation of our solutions in Google Custom Search Engine as well as with the automatic acquisition of the textual contents from HTML code. Web pages do vary widely, so it was necessary to develop customized text parsers to enhance the acquisition of the content. We experienced that many web pages contain errors in the HTML code, which may cause some problems in obtaining web content efficiently, so it was necessary to develop a solution that supports an option of manual input or editing by the user. One of the greatest challenges was also how to generate screenshots and store them into the database. Obtained screenshots show real content of web pages with high resolution, which consequently takes a lot of space. There were some problems in finding the optimal format of screenshot that takes minimal space and still supports satisfactory resolution.

With explosion of data available on the web people have access to stuff more effectively across the globe, being able to get information which they haven’t been able to get before. Technological progress is remarkable. From year to year faster and more powerful computers are built. We can use their speed and power into filtering data, and then allow them to join that knowledge to other summary data to maximize the human’s ability to see the correlation of the trends. It may be that the sea change will be greater; maybe we will find out some now patterns with the sea of data and be able to solve problems that we haven’t solve before.

4 Acknowledgements

"Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund and Nevtron & Company d.o.o. research voucher founded by Ministry of Education, Science and Sport, Slovenia. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and re-search excellence."

5 References

[1]Alexa, http://www.alexa.com/, downloaded: April 24th 2014.
[2] Cortes, C.; Vapnik, V. Support vector networks. Machine learning, 20:273-297, 1995.
[3] Google Alerts, http://www.google.com/alerts, downloaded: May 5th 2014.
[4] Google Custom Search Engine, https://www.google.com/cse/, downloaded: April 15th 2014.
[5] Lemmatise, http://lemmatise.ijs.si/Software/, downloaded: May 20th 2014.
[6] Lewis, D.D. Naïve (Bayes) at Forty: The Independent Assumption in Information Retrieval. In Proceedings of a 10th European Conference on Machine Learning, pages 4-15, Chemnitz, Germany, 1998.
[7] Likert, R. A Technique for the Measurement of Attitudes. Archives of Psychology, 22: 1-55, 1932.
[8] Liu, B. Sentiment analysis and subjectivity. Handbook of natural language processing, 2: 627-666, 2010.
[9] Miner, G.. Practical text mining and statistical analysis for non-structured data applications. Amsterdam: Elsevier, 2012.
[10] PhantomJS – Screen Capture, http://phantomjs.org/screen-capture.html, downloaded: April 26th 2014.
[11] Popescu, A.; Etzioni, O. Extracting product features and opinions from reviews. Natural language processing and text mining, Springer, London, Great Britain, 2007.
[12] Sanger, J.; Feldman, R. The Text Mining Handbook. Cambridge Univ. Press, New York, 2007.
[13] Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of an Association for Computational Linguistics, pages 417-424, Pennsylvania, USA, 2002.
[14] Web Clipping, http://www.Webclipping.com, downloaded: April 10th 2014.
[15] Wright, A. Mining the Web for Feelings, Not Facts. New York Times, 2009.

Information System Mirror – Approach How to Analyze Information System Strengths and Weaknesses within Organization

Boštjan Delak
ITAD, Revizija in svetovanje, d.o.o. Technology Park Ljubljana
Pot za Brdom 100, 1000 Ljubljana, Slovenia, www.itad.si
bostjan.delak@itad.si

Abstract: *Nowadays, many organizations face the issue of information and communication technology (ICT) management. Companies undertake various activities to assess the state of their ICT or the entire information system (IS), such as IS audits, reviews, IS due diligence, or they already have implemented systems to gather information regarding their IS. However, all the above activities take quite some time to conduct interviews, testing and verification of sources for implementing various systems. The IS Mirror approach is fast and simple method to analyze the efficiency of IS management, which can be done remotely with an online questionnaire. After the answers have been completed, they are analyzed and a report with recommendations for improvements is generated for the client. Over the past few months, the analysis has been conducted in several Slovenian organizations from different sectors of the economy. The paper presents the cumulative results of the analyses performed so far.*

Key Words*: Information System, IS Analysis, IS strengths and weaknesses, IS Mirror*

1 Introduction

Information and communication technology (ICT) is included in most of the processes of the organization and therefore it is extremely important to manage this area. Information systems (IS) are much more than ICT, including the processes, data, documentation, and people - ICT professionals and end users. ICT professional's daily care that IS in the organization functions smoothly while maintaining and upgrading it both in terms of hardware and system software as well as application software. Stakeholders, management and owners often wonder whether their IS have sufficient quality and efficiency to meet the objectives, allow end user quality support for their daily operation and provide the management with sufficient information to make the right decisions. Usually, management and owners obtain a variety of information about the state of the IS by both ICT professionals and from end users.

In order to obtain an independent report about the status of their IS, some organizations perform different activities: IS audit, other audit and even IS due diligence. Each of these activities also identifies the presence or lack of certain controls, non-compliance or other findings. Some organizations have already introduced systems that they can answer to the above question. These systems are: a quality management system [8], the system of IT service management [9], Balanced Scorecard [12, 2], and others. Another possible answer is to carry out IS due diligence. However, each of these systems requires an already implemented system or long and time-intensive due diligence

review. Our position is that it is possible to make a first assessment in an easy and fast way using the IS Mirror method in a similar way as the methods for measuring human competencies (for example, 360-degree feedback [14]). The IS mirror method is detailed described within this paper.
The paper has eight chapters. In the first part we present the IS analysis through the IS due diligence. We continue with the presentation of our motivation and hypotheses presentation. The fourth part describes some previous experiences written in papers in Slovenia and abroad. The fifth part describes in detail the IS mirror method, the way of performance and structure of the report for the client. In the sixth section we present lessons learned and some of the results. Followed by a discussion, in conclusion, we present opportunities for further activities and further development of the method.

2 IS due diligence

To obtain the current IS state within the organization and consequently the degree of IS presence in the organization carried out a careful examination of IS. Due diligence terminology is explained in Wikipedia [19] as.

> *"Due diligence is an investigation of a business or person prior to signing a contract, or an act with a certain standard of care".*

The basic aim of the IS due diligence is to obtain information about the quality and performance of the IS, their sources, documents, risks and processes. It is necessary to obtain information on the status and effectiveness of the internal control system (quality control environment of functioning), to obtain information about the risks in the field of IS, the pros and cons of IS and the rate of coverage of products / services / processes IS. Delak and Bajec [3] cite a number of ways due diligence. All of these methods have much in common, but differ primarily according to whom the customer is and what are the objectives of the review. In Slovenia we have developed a Framework for IS due diligence (FISDD) [3], which consists of four phases: preparation, an on-site review, analysis and decision-making. As part of a comprehensive approach for the on-site review several questionnaires have been prepared. One of them is also IS strengths and weaknesses questionnaire.

3 Motivations

Different types of audits and also miscellaneous IS due diligence required human resources and the presence of experts at the client's location. With the objectives to carry out a brief but effective IS analysis of the inspected organization, we have prepared a base on our Framework for IS due diligence (FISDD) special method for such an activity entitled it IS mirror method. Our hypothesis is that:
"*With the IS mirror method is possible to quickly and efficiently gather enough data for an analysis of the strengths and weaknesses of IS in the organization and on the basis of experience to draw up recommendations for improvement, which allow the subscriber appropriate action".*

4 Literature review

In the world there are no standard guidelines for the implementation of IS due diligence activities. There are quite a few ways, standards, methodologies and best practices to perform these tasks. Delak and Bajec [3] indicate possible ways of implementation due diligence. On the other hand, there have been carried out various analyses of IS implementation. Damij and Gebert [2] had analyzed several Slovenian companies, which had introduced ERP (Enterprise Resource Planning) systems, and argue that the implementation of ERP systems concerns and improve the company acceded to the

restructuring of business processes. Erjavec et al. [6] showed the state of business informatics in Slovenia, and present a noticeable change in the field: the organization of the IS departments and needs different tools for data management. Erjavec et al. like Damij and Gebert noted that business processes are becoming important, but so far their managements have not yet been efficiently realized. Vrhovec and Rupnik [18] discussed how to cope with changes in the IS and thus, consequently, how to manage resistance of IS end users by generic model. They have pointed out that the evolution of end users is vital for the progress of the organization and also their vital role in the communication. Cerovšek [1] pointed out that the search of effectiveness levers and efficiency of management structure's role in the organization, which can be realized by the growing importance of ICT and focused on the ICT usage, business process management and integration and development of employees. Cerovšek also emphasized keeping proper communication between all involved structures.

On the other hand, by the review of certain scientific papers we observed descriptions of the various methods for evaluating the effectiveness, quality and benefits of IS. Sedera and Tan [17] pointed out that the satisfaction of end users is the most widely used dimension to ensure the success of IS and its contribution to analyze factors: the quality of information, quality system, the impact on the individual and the impact on the organization. Zviran and Erlich [20] emphasized that measuring the success of IS increases with the importance and involvement of IS in the organization. At the end of the 70s and the beginning of the 80s of last century, it was examined end user satisfaction with their IS by the UIS (*User Information Satisfaction*) method indicated by Ives et al. [10]. In the second half of the 80s TAM (*Technology Acceptance Model*) was developed by Davis and eventually became the most used theoretical models in the field of IS [15]. DeLone and McLean [4] developed a *DeLone and McLean model* (*D&M model*) of IS success based on six dimensions: system quality, information quality, users, user satisfaction, individual impact and the impact on the organization. This model was supplemented over the years and also compared with TAM [7]. The authors of D&M model have upgraded a model by the dimension of e-commerce a decade after the emergence [5]. Rabaa'i [16] described a comparison of the D&M model with Gablov's model. Rabaa'i has focused on two areas of the respondents: management and end users. In addition to the D&M model for the assessment of the SERVQUAL model occurs [13], which assessed the IS through quality service by scoring of five different dimensions. Jia et al. [11] developed a model that builds on SERVQUAL model from a psychological point of view of climate in the organization through IT service climate, which is four-dimensional and contains areas: service management, vision of the services, end users feedback and communication with end users.

5 IS Mirror method

A framework for IS due diligence (FISDD) contains a questionnaire to gather information on the pros and cons of IS. This questionnaire and our experience with it led us to the preparation of the IS Mirror method that in times of economic crisis allows the IS self-assessment (mirror) in the organization. IS Mirror is fast and simple method for analysis of the effectiveness of IS management. Based on the COBIT methodology, it is complemented by soft factors of cooperation between the end users and ICT.

5.1 Method's benefits

The IS Mirror method:

- Provides an insight into the effectiveness of ICT support;
- Measure the quality of cooperation between the end users and ICT;

- Discovers hidden reserves to increase the efficiency and effectiveness of IS;
- Recommends measures to increase the efficiency of the IS.

The method allows the rapid and efficient implementation. From the first contract to the report, could be made in two (2) weeks. On average, the execution time is between three to four weeks.

5.2 Method's phases

IS mirror is composed of four phases. Graphical representation is shown in Figure 1. At the beginning, it is very important to choose survey's participants. It is important that both the end user as well ICT include different profiles of respondents. Required are their e-mail addresses and indication whether the participant is the end user or ICT.

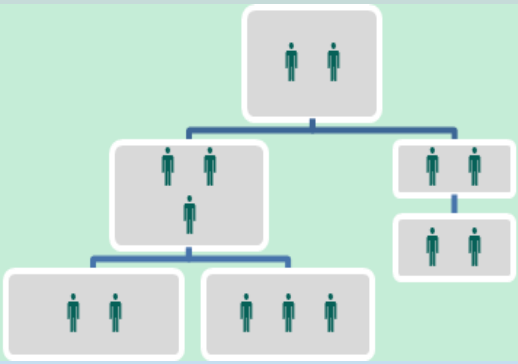
Phase		Number of participants	Duration
1	Selection of participants in the survey		2 hours
2	The survey participants		30 minutes each
3	Analysis of the survey and report set up		3 - 4 hours
4	Presentation of findings and recommendations, handing over a written report		1 – 2 hours

Figure 1: Mirror IS – presentation of the individual phases

In the second phase, the selected participants receive an email and a link to the online questionnaire. It consists of 58 questions within 8 groups: Productivity of the IS, Quality of the existing application system, Effective use of technology, Information security (confidentiality, integrity, availability information), Usage the advanced and modern technologies, ICT employees, Cooperation between end users and ICT, Participation of ICT in projects. Each respondent first assesses whether the item to which the question refers is an advantage (+) or a disadvantage (-). Then give a numerical value, which can be for "strength" from +5 (ideal / can not be better) to +1 as the minimum strength. If the item is in a "weaker" estimates can be from 1 (minimal weakness) to -5 (worst / can not be worse). If the respondent does not have the experience or cannot answer a question, then give the mark 0 and comments it. The respondent can enter the reason for giving mark (for example: I cannot identify the answer, I do not have experience, I do not know the area). Some questions also have a smaller range of possible numerical answers. For each question there is also available help, assistance in order to provide more detailed explanation about the question.

According to our estimates and experience the time to respond to a survey is about 30 minutes.
The third phase begins when all participants respond. Data is exported and processed in order to obtain the results. IS Mirror method can also From the analysis independently assess the capability maturity model (CMM) of the IS support. CMM has six levels: 0 - No, 1 - Initial / Ad Hoc, 2 - Repeatable but intuitive, 3 - Defined, 4 - Managed and measurable, 5 - Optimized. Specialist prepares the report that consists of three chapters: Report's summary; Details, which consists of: purpose, scope and content, statistic's results, details of particular areas, assessment and restrictions; and recommendations. Normally report comprises between 12 and 15 pages. The results of the analysis are presented in tables, graph (Figure 2 - Example of graph for a client from electric industry), as well as descriptive.
Phase four is when the client requests a short meeting to present the results of the analysis with recommendations.

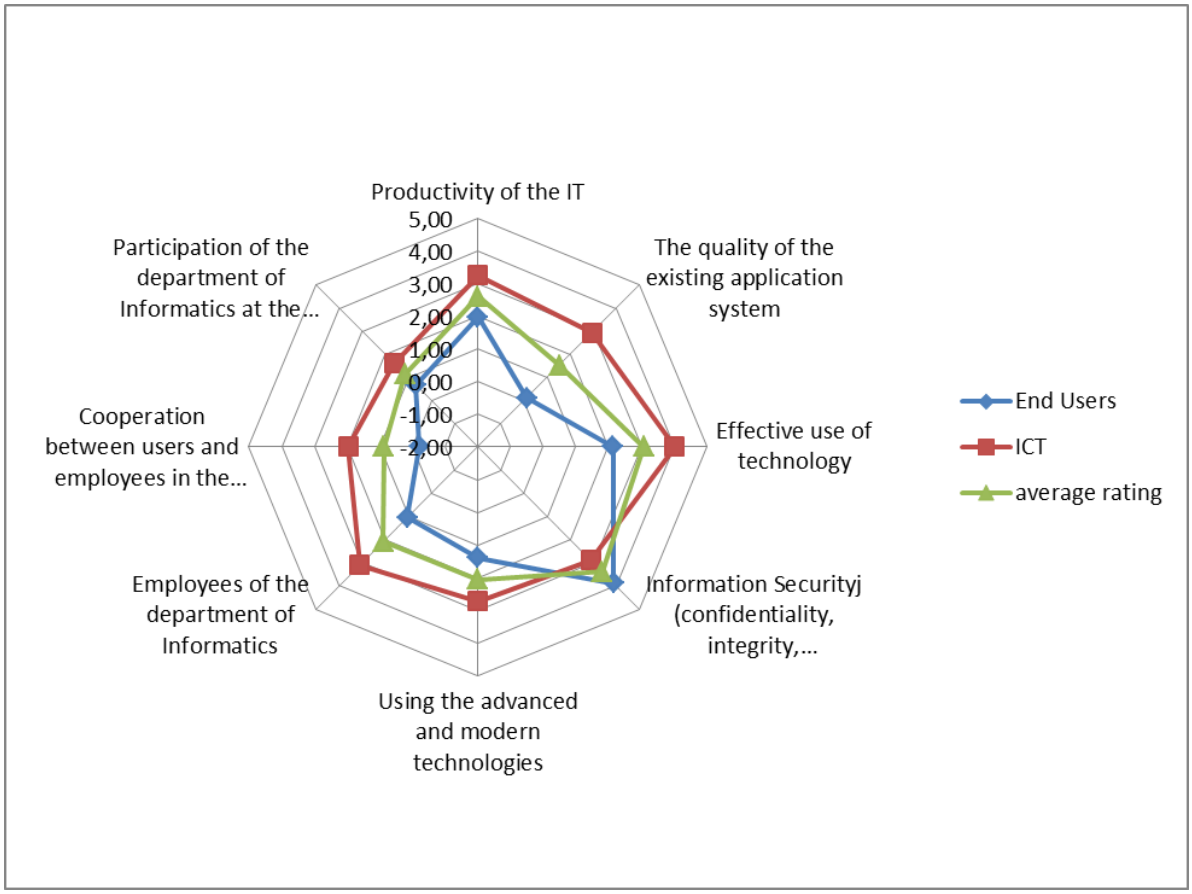


Figure 2: IS Mirror - a graphical representation of responses by area

6 Lessons

In recent months, we performed several analyzes in four different branches of economy. Table 1 presents the gathered data from all different branches. The findings are as follows:

- Most of the ICT answers are more positive than from the end users;
- End users assess the quality of the existing IS more realistic (worst rating) that ICT;
- ICT give more realistic marks in the field of information security;
- CMM assessments were realistic (so far we have had marks from 1 to 3).

Table 1: Numeric results for four different industries

	Electricity			Institute			Manufacturing			Government		
description	E.U.	ICT	Ave.	E.U.	ICT	Ave.	E.U.	ICT	Ave.	E.U.	ICT	Ave.
Productivity of the IT	1,97	3,25	2,61	0,84	-0,67	0,08	2,23	2,73	2,48	2,81	1,00	1,91
The quality of the existing application system	0,12	2,93	1,53	0,26	0,36	0,31	2,20	2,66	2,43	3,13	3,82	3,48
Effective use of technology	2,14	4,02	3,08	1,57	4,33	2,95	2,92	2,30	2,61	2,47	2,50	2,49
Information Securityj (confidentiality, integrity, availability information)	3,89	2,89	3,39	1,39	2,22	1,81	3,17	3,87	3,52	3,60	3,00	3,30
Using the advanced and modern technologies	1,39	2,72	2,06	-1,16	-2,83	-2,00	0,44	3,26	1,85	1,27	-1,08	0,10
Employees of the department of Informatics	1,05	3,11	2,08	0,83	-2,17	-0,67	2,22	2,25	2,23	3,33	3,83	3,58
Cooperation between users and employees in the department of Informatics	-0,21	1,96	0,87	0,26	-0,56	-0,15	1,98	2,46	2,22	2,41	2,83	2,62
Participation of the department of Informatics at the projects	0,68	1,61	1,15	0,03	1,58	0,81	2,29	2,73	2,51	3,21	2,50	2,86
Average	**1,38**	**2,81**	**2,09**	**0,50**	**0,28**	**0,39**	**2,18**	**2,78**	**2,48**	**2,78**	**2,30**	**2,54**

Some clients have already expressed their desire and plans to periodically (every year or every two years) repeat the IS Mirror analysis and monitor the progress.
With these real case studies that we have carried out in recent months by different organizations, we got the confirmation of our hypothesis *that it is possible with the method of mirror IS quickly and efficiently gather enough data to be able to undertake an analysis of the strengths and weaknesses of IS in the organization and based on experience, make recommendations for improvement, which allow the subscriber appropriate action*.

7 Discussion

Measurement of IS: performance, quality and satisfaction in practice and in science has been practiced for decades. In the scientific field, this is one of the most researched and outlined areas with many papers describing and analyzing methods, approaches and case studies. Delak and Bajec [3] in the description of the framework for IS due diligence emphasized that the approach is a set of positive parts of different methods, approaches, best practices and standards. This also applies to the scope of the IS strengths and weaknesses questionnaire, which indirectly covers the areas covered by the D & M model describing the Delone and McLean [5] and Jia et al. [11] with the orientation of the model, IT service climate. We believe that similar parallels can be found in other methods for measuring performance, quality and satisfaction of IS in comparison with the described IS Mirror method. It should be emphasized that the IS mirror built on the basis of a questionnaire, which was created through a multi-year delivery of IS due diligence in various countries of Central Europe [3].

8 Limitation

Our research has several limitations: the approach has been tested only in Slovenia and in small enterprises. As the approach has been developed from the framework for IS due diligence it has not yet been compared with the results from other approaches for measuring the efficiency of IS management.

9 Conclusion

This paper describes a method for fast and effective analysis, reviews the operational efficiency of IS in the organization. The method is carried out remotely via electronic communications (online questionnaire) and is based on the analysis of the pooled estimate advantages (+) and cons (-) made by the different profiles of end users and ICT. The analysis of the collected data it is possible to assess the CMM of IS in the

organization. The paper presents a hands-on experience, as was the IS mirror method carried out in four different branches of the economy, thus confirming the possibility to use it in any organization regardless of size and activity in Slovenia. Our plans for upgrading methods related to the integration of the evaluation of soft skills of end users and ICT in the organization. On the other hand, we plan to internationalize methods in terms of translation into the English language and to the languages of neighboring countries of Slovenia and the penetration of this method in the space of Central Europe. In the long term we plan various analyzes and studies of collecting data and preparation of different papers at Slovenian and international conferences.

9 References

[1] Cerovšek, M. Informatika mora pokazati svojo poslovno vrednost, Uporabna informatika, No.3:195-201, 2012.

[2] Damij, N., Gebert, A. Slovene Companies and ERP Systems, Proceedings: ISIT2010 – 2nd Conference on Information Society and Information Technologies, 2010.

[3] Delak, B., Bajec, M. Framework for Information System Due Diligence, Information System Management, 30 (2): 137-149, 2013

[4] DeLone, W., McLean, E. Information System Success: The Quest for the Dependent Variable, Information System Research, 3 (1): 60-95, 1992.

[5] DeLone, W., McLean, E. The DeLone and McLean model of Information System Success: a Ten-year Update, Journal of Management Information System, 19 (4): 9-30, 2003.

[6] Erjavec, J., Groznik, A., Gradišar, M., Indihar Štemberger, M., Jaklič, J., Kovačič, A., Turk, T., Popovič, A., Trkman, P., Manfreda, A. Analiza stanja poslovne informatike v slovenskih podjetjih in javnih organizacijah, Uporabna informatika, No.1:44-51, 2010.

[7] Hellsten, S-M., Markova, M. The DeLone and McLean Model of Information Systems Success – Original and Updated Models, http://www.cs.tut.fi/~ihtesem/s2006/teoriat/esitykset/IS_success_model_Markova&Hellsten_311006.pdf downloaded: January 24th 2014.

[8] ISO. ISO 9001:2005 - Quality Management Systems, 2005.

[9] ISO/IEC. ISO/IEC 20000 – IT Service management system, 2011.

[10] Ives, B., Olson, M.H., Baroudi, J.J. The Measurement of User Information Satisfaction, Communication of the ACM, 26 (10): 785-793, 1983.

[11] Jia, R., Reich, B.H., Pearson, J.M. IT Service Climate: An Extension to IT Service Quality Research, Journal of the Association for Information Systems, 8 (5): 294-320, 2008.

[12] Kaplan, R.S., Norton, D.P. The balanced scorecard: translating strategy into action, Harvard Business Press, 1996.

[13] Landrum, H., Prybutok, V., Zhang, X., Peak, D. Measuring IS System Service Quality with SERVQUAL: Users' Perceptions of Relative Importance of the Five SERVPERF Dimensions, C Informing Science: the International Journal of an Emerging Transdiscipline, 12, 2009.

[14] London, M., Beatty, R.W. 360-Degree Feedback as a Competitive Advantage, Human Resource Management, Vol.32, No. 2 & 3: 357-372, 1993

[15] Lee, Y., Kozar, K.A:, Larsen, K.R.T. The Technology Acceptance Model: Past, Present, and Future, Communications of the Association for Information Systems, 12 (50): 752-780, 2003.

[16] Rabaa'i, A.A. Assessing Information System Success Models: Empirical Comparison (Research in Progress), Proceedings: 20th Australian Conference on Information Systems, Paper 61, 2009.

[17] Sedera, D., Tan, F. User Satisfaction: An Overarching Measure of Enterprise System Success, PACIS 2005 Proceedings, Paper 80, 2005.

[18] Vrhovec, S., Rupnik, R. (2011). Obvladovanje odpora pri projektih informacijskih tehnologij, Uporabna informatika, No. 1:23-37, 2011.

[19] Wikipedia. (2014) item = due diligence Available at: http://en.wikipedia.org/w/index.php?title=Special:CiteThisPage&page=Due_diligence&id=630590285, accessed: June 11th 2014.

[20] Zviran, M., Erlich, Z. Measuring IS User Satisfaction: Review and Implications, Communications of the Association for Information Systems, No.12, Paper 5:81-103, 2003.

Bottleneck is on the top of the bottle or how to improve throughput in IT development

Tomaž Aljaž
Faculty of Information Studies
University of Novo mesto
tomaz.aljaz@gmail.com

Abstract: *The goal of every organization is to have satisfied owners, their clients and own employees today and in the future. If any of these are missing, the organization needs to address the undesirable effects that are causing them to be present and preventing organization to achieve its goal. The results presented in the document are focused on the tools and techniques provided by Theory of Constrains in order to improve performance of the organization. The "Theory of Constrains - TOC" is described as set of tools and techniques that are aimed to bring improvement in the overall performance of organizations by focusing on a few important points.*
In almost any organizations, there are plenty of actions that are expected to contribute to the overall performance, but usually there is not enough time, money or resources. The TOC brings approach that enables recognizing few important from many trivial points that organization needs to address in order to achieve competitive advantage. Important points that limit organization from achieving higher performance towards its goal (competitive advantage) are called constrains or bottlenecks. Recognizing who or what is organization's constrain enable us to focus on the core of the points and not on their symptoms.
In the presentation we will show how five focusing steps defined by TOC can be applied in IT development area and improve throughput of deliverables. The five focusing step are defined as: (1) identify constraint, (2) exploit the constraint, (3) subordinate the constraint, (4) elevate the constraint and (5) Warning ; if constraint moves, go back to step 1. Especially, in the presentation we will discuss different types of constraints and how we are able to manage them, followed by the solution to overcome current reality and improve overall results.
Finally, having in hand these powerful tools enable us (management) to focus on few important points and not on their symptoms. With clear indication what needs to be improved, it defines and rearranges existing management priorities in order to increase the quality of deliverables, establish long term cooperation with the clients / users and to improve our internal cooperation. There is win-win solution for owners, organization's clients and employees.

Key Words*: Information technology, development, Theory of Constraints, Agile development…*

Diagnosing mental disorders as a result of changes in the autonomic nervous system function

Albert Zorko, doc. dr. Zoran Levnajić
Faculty of Information Studies
University of Novo mesto
Sevno 13, 8000 Novo mesto, Slovenia
albert.zorko@gmail.com
zoran.levnajic@fis.unm.si

Abstract: *Psychological states have a huge impact on autonomic nervous system. The literature on this science area is far from simple, however mental states effect on many states of autonomic nervous system. In following paper we highlight an overview of important researches on this area. We were considering different conceptual approaches which help us to understand better from different perspective this interesting and for human health very important scientific area.*

Key Words*: mental disorders, heart rate variable, diagnostics, cardiorespiratory coupling*

1 Introduction

Background: A large amount of clinical psychophysiology researches deals with autonomic dysfunctions. Methods investigate instead of all autonomic nervous system only a subordinate autonomic structure. In order to achieve a new breakthrough is necessary to introduce new methods.

Deferent methods are used by researcher to evaluate autonomic dysfunction. Very commonly used method is measuring heart rate and respiratory rate variability. Frequency specific fluctuation of heart rate is assessed with power spectrum analysis. Other used methods are: electrocardiography, electroencephalography, blood pressure variability measure, microneurography, mean arterial pressure and transcutaneous oxygen.

1.1 Autonomic nervous system disorders

Autonomic nervous system is part of human nervous system. It is consist of all nervous pathways leaving central nervous system that have a ganglionic synapse situated outside the central nervous system. There are three distinct anatomical divisions: sympathetic, parasympathetic and enteric nervous system [1]. Functioning of autonomic nervous system is strongly depends on equilibrium between parasympathetic and sympathetic nervous system. Complex structure of autonomic nervous system can be seen in Figure 1.

Parasympathetic activation results in slowing of the heart rate, a reduction in the force of contraction in the atria and a reduction in conduction velocity through the atrioventricular node [1].

Sympathetic activation results in increases the force and rate of cardiac contraction, blood pressure increases from both increased peripheral resistance and increased cardiac output, increasing salivation, dilates the pupils in the eye, increases sweating, evokes piloerection, inhibits sodium excretion, and causes trembling [1].

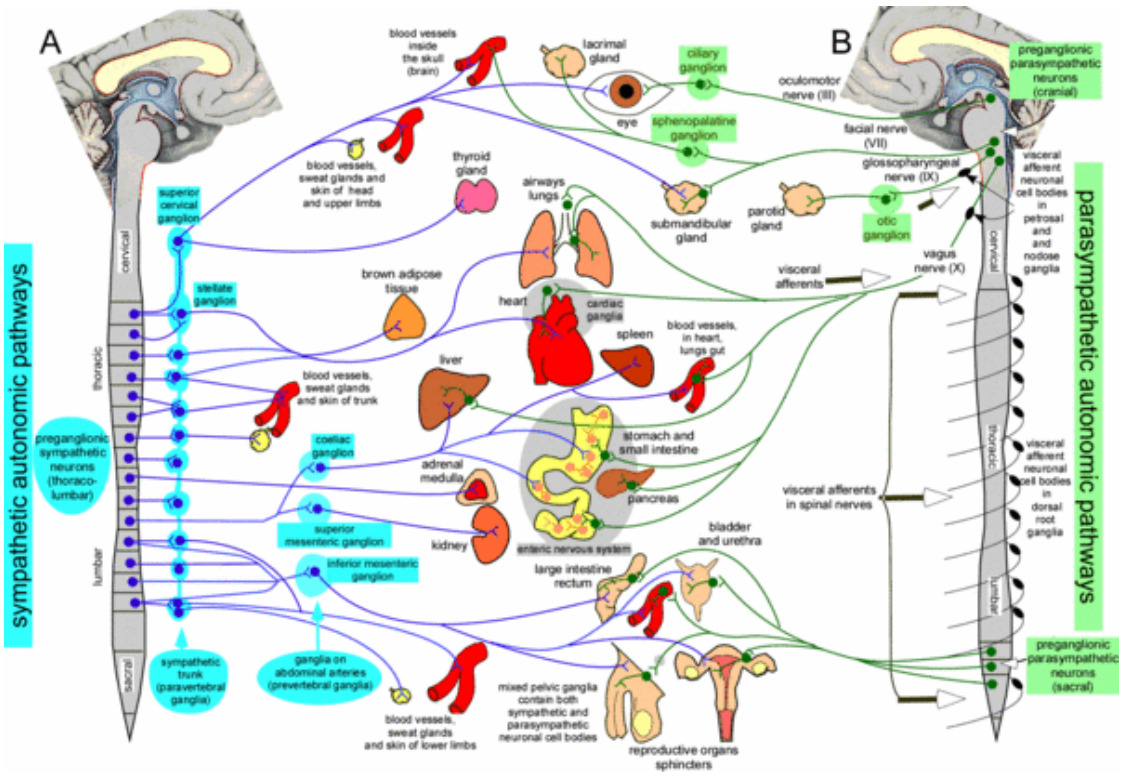


Figure 1: Autonomic nervous system [2]

Many factors influence from on the autonomic nervous system, including mental state. As an example we should look at Anxiety, Obsessive Compulsive Disorder and Post Traumatic Stress Disorder. Real cause is steel unknown, but there are many known triggers for these mental disorders. In addition to external factors chemical imbalances in the nervous system are very common.

Typical example is excess adrenaline production. Glucose is the forerunner of Biological Energy called (ATP), which is essential in the manufacture of the relaxing and feel good neurotransmitters, such as serotonin. When hypothalamic–pituitary–adrenal axis (HPA axis) senses a low blood sugar level it will send a hormonal message to the adrenal glands to pour adrenaline into the system. This raises blood sugar level and will feed the brain again, but it also causes us to feel fearful without an external object of fear. There are many reasons for this, because there are many medical conditions that interfere with the proper absorption of glucose. [3].

Another example is bipolar disorder (manic - depressive illness). The triggers can be all sorts of life events with positive or negative nature. Illness is associated with changes in various neurotransmitter levels and activity, commonly referred to as a chemical imbalance in the brain [4].

2 Diagnostic methods

2.1 Classical diagnosis

Classical diagnosis began with physician asking questions about medical history and medical history and sometimes physical exam. There are no laboratory tests to specifically diagnose mental illness. Doctor uses different tests to exclude other diseases. A specially designed interview and assessment tools are used to evaluate a person for mental illness [5]. World Health Organization has approved two questionnaires GHQ - 12[1] and GHQ - 28. Results were uniformly good when using both questionnaires [6]. Diagnosis become very difficult when differential[2] diagnosis must be made.

2.2 Magnetic resonance imaging

Researchers are seeking for alternate diagnostic methods, for more objective diagnosis. Magnetic resonance imaging (MRI) is promising method, but not for all mental disorders. Better accuracy is achieved for ADHD[3], schizophrenia, Tourette's and bipolar disorder. More research still needs to be done to ensure the success [7].

Smaller hippocampal volume predicts pathologic vulnerability to severe stress [8]. Douglass Bremner with his associates using MRI discovered that patients with depression had a statistically significant 19% smaller left hippocampal volume than comparison subjects, without smaller volumes of comparison regions (amygdala, caudate, frontal lobe, and temporal lobe) or whole brain volume. The findings were significant after brain size, alcohol exposure, age, and education were controlled for [9]. Some other caveats are: system will probably struggle with patients who have more than one diagnosis, system is unable to detect early stages of disorders and diagnostic categories might not be biologically valid [10].

2.3 Electroencephalography (EEG)

Statistical machine learning methodology on EEG data is used for diagnosis of psychiatric illnesses: major depressive disorder, chronic schizophrenia and bipolar depression. The average correct diagnosis rate attained using the proposed method is over 85%, as determined by various cross – validation experiments [11]. A novel research ideology, multi-paradigm methodology and advanced computational models for automated electroencephalogram (EEG) – based diagnosis of neurological and psychiatric disorders was presented on international conference. Model should be used for automated diagnosis of epilepsy, the Alzheimer's disease, Attention Deficit Hyperactivity Disorder (ADHD), and Autism Spectrum Disorder (ASD) [12].

[1] General Health Questionnaire
[2] diagnosis for multiple illness
[3] Attention-Deficit/Hyperactivity Disorder

2.4 Heart rate variable (HRV)

Well established method for the diagnosis of mental disorders is the measurement of heart rate variability. It can be measured by variation in the R-R interval of electrocardiogram (ECG). For further reading we suggest reading [13-14]. Billman has described a historical perspective of heart rate variability. So, further reading in paper [15] is recommended. H. Tsuji study showed that reduced hart rate variability has been reported to predict risk for subsequent mortality [16-17].

Gary G. Berntson and John T. Cacioppo showed that stress has clear impact on autonomic nervous system. It can be seen through lowering the HRV [18]. Netherland study shows that depression is associated with significantly lowered heart rate and respiratory rate variability. However, this association appears to be mainly driven by the effect of antidepressants [19]. Andrew H. Kemp with associates study compares HRV in patients with major depressive disorder and healthy control subjects and the HRV of patients with major depressive disorder before and after treatment was considered for meta-analysis. They conclude that depression without cardio vascular disease is associated with reduced HRV, which decreases with increasing depression severity, most apparent with nonlinear measures of HRV. Critically, a variety of antidepressant treatments (other than TCA's[4]) neither increases nor decreases HRV [20]. Another study concluded that adolescent female psychiatric patients with anxiety disorder and/or major depression disorder had a reduced HRV compared with healthy controls. Medication with SSRI[5] explained a part of this difference [21]. Chalmers with his team found that anxiety disorders are associated with reduced HRV; findings are associated with a small-to-moderate effect size. Findings have important implications for future physical health and well-being of patients, highlighting a need for comprehensive cardiovascular risk reduction [22]. Similar research findings are presented by [23-25]. One of plausible prevention suggest that modest amounts of regular moderate-to-vigorous physical activity sufficient to slow the accelerating age-related decline in cardiorespiratory fitness during late middle-age has protective benefits against the onset of depression complaints in both men and women [26]. Gorman and Sloan published on observation of heart rate variability that depressed patients after myocardial infarction, exhibit higher mortality rates compared with non-depressed patients. Men with "phobic anxiety," also have higher rates of sudden cardiac death and coronary artery disease than control populations. Although HRV is reduced, treatment with the selective serotonin reuptake inhibitor paroxetine normalizes heart rate variability. Hence there is potential for the treatment of psychiatric disorders to affect positively the development and course of cardiovascular disease [27].

2.5 Pulse – respiratory coupling

Some studies observe pulse in respiratory system separately. They analyzed ECG and respiratory signal and search for possible connections with mental disorders [19]. Widjaja with his college found that during stress and attention heart and respiratory rate is increased in compared to a resting condition [28].

A novel approach has been performed; instead of separately analyzing cardio and

[4] Tricyclic antidepressants (TCA's) are chemical compounds used primarily as antidepressants.
[5] selective serotonin reuptake inhibitor is used as antidepressant

respiratory data, observation of autonomic nervous system through cardiorespiratory coupling is used. Kralemann and his colleges show that the phase at which the cardiac beat is susceptible to respiratory drive and extract the respiratory-related component of heart rate variability [29]. Möser and his team use a multidimensional approach to describe autonomic nervous system and conclude that only with peripheral autonomic activity measurement we can observe a central events influence of this activity in coordination of physiological parameters [30]. There are still many ongoing researches on this field and we can expect a breakthrough in this field of research.

3 Conclusion

Serious work has been done in area of observation of autonomic nervous system. Researches describe many influence factors that can interferes function of the ANS. Among of them are mental disorders which take serious proportion in human population. In future study a systematical investigation in lager group of subject and patients must take place, so we once can have automatic, no subjective method to diagnose mental disorders. Huge gap is still to overcome in area of multiple mental disease diagnostics.

In future, we expect that the diagnostics will rely more on the analysis of the data. In fact, modern biomedical experiments generate enormous amounts of data on various body parameters, among them cardio-respiratory data and HRV data. Using statistical analysis, data mining methods, automated machine learning procedures we could be able to have faster and more accurate prediction of diagnosis and hopefully have some new progress in differential diagnosis. Those findings may find an important place in medical practitioner toolbox. It is particularly important for initiate appropriate treatment when expert psychiatric assessment may not be available for many weeks.

4 Acknowledgments

We would like to thank Max Möser for consultation and Faculty for information studies for support. We also thank Ministry of education, science and sport, Republic of Slovenia for financing.

5 References

[1] Brading, A. Autonomic Nervous System and Its Effectors, Wiley-Blackwell, Hoboken, New Jersy, 1999.
[2] Blessingm, B; Gibbins, I. Scholarpedia. Autonomic nervous system, http://www.scholarpedia.org/article/Autonomic_nervous_system, (accessed: Oct 9, 2014).
[3] Plesman, J. Anxiety and the Autonomic Nervous System, http://www.hypoglycemia.asn.au/2011/anxiety-and-the-autonomic-nervous-system/, (accesses: Oct 12, 2014).
[4] Albrecht, AT.; Herrick, C. 100 questions and answers about bipolar (manic

depressive) disorder. Jones and Bartlett Publishers, Sudbury, Massachusetts, 2007.
[5] Goldberg, J. Diagnosing Mental Illness. http://www.webmd.com/anxiety-panic/guide/mental-health-making-diagnosis, (accessed: Oct 10, 2014).
[6] Goldberg, DP et al. Psychological Medicine. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Cambridge Journals Online, 27(1), 1997.
[7] Marlow, K. Diagnostic Breakthrough for Mental Illness, http://www.psychologytoday.com/blog/the-superhuman-mind/201212/diagnostic-breakthrough-mental-illness, (accessed: Oct 13, 2014).
[8] Gilbertson, MW. Smaller hippocampal volume predicts pathologic vulnerability to psychological trauma. http://www.ncbi.nlm.nih.gov/pubmed/12379862, (accessed: Oct 15, 2014).
[9] Bremner, JD et al. Hippocampal Volume Reduction in Major Depression. http://journals.psychiatryonline.org/article.aspx?articleid=173909, (accessed: Oct 15, 2014).
[10] Makin, S. Scientific American. Can Brain Scans Diagnose Mental Illness?, 24(3), 2013.
[11] Khodayari – Rostamabad A. et al. Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model, EMBS Annual International Conference of the IEEE, Buenos Aires, Argentina, 2010.
[12] Adeli, H. Automated EEG-based diagnosis of the neurological and psychiatric disorders. 18th International Conference on Systems, Signals and Image Processing (IWSSIP), 2011 Sarajevo, Bosnia and Herzegovina, 2011.
[13] Berntson, GG et al. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology, 34(6), 623 – 648, 1997.
[14] Miyamoto, M; Ichimaru, Y; Katayama, S. Heart rate variability. Nihon rinsho, 50(4): 717 – 22, 1992.
[15] Bilmann, GE. Heart rate variability - a historical perspective. Frontiers in psysiology. November, 2011,
[16] Tsuji, H et al. Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study. Circulation. 90(2): 878 – 883, 1994.
[17] Tsuji, H et al. Impact of reduced heart variability on risk of cardiac events The Framingham Heart Study. Circulation. 11(94): 2850 – 2855, 1994.
[18] Berntson, GG; Cacioppo, JT. Dynamic electrocardiography. Futura publishing company Inc., New York, USA, 2004.
[19] Licht, CMM et al. Association between major depressive disorder and heart rate variability in the Netherlands Study of Depression and Anxiety (NESDA). Archives of General Psychiatry. 65(12): 1358 – 1367, 2008.
[20] Kemp, AH et al. Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. Biological Psychiatry. 67(11): 1067 – 1074, 2010.
[21] Blom, HE et al. Heart rate variability (HRV) in adolescent females with anxiety disorders and major depressive disorder. Acta Paediatrica. 99(4), 2010.
[22] Chalmers, JA et al. Anxiety Disorders are Associated with Reduced Heart Rate Variability: A Meta-Analysis. Front Psychiatry. 11(5), 2014.
[23] Miu, AC. et al. Reduced heart rate variability and vagal tone in anxiety: trait versus state, and the effects of autogenic training. Autonomic Neuroscience: Basic and Clinical. 145(1–2): 99 – 103, 2009.
[24] Friedman, BH; Thayer, JF. Autonomic balance revisited: Panic anxiety and heart rate variability. Journal of Psychosomatic Research. 44(1):133 – 151, 1998.
[25] Thayer, JF; Friedman, BH; Borkovec TD. Autonomic characteristics of generalized

anxiety disorder and worry. Biological Psychiatry. 39(4): 255 – 266, 1996.
[26] Dishman, RK et al. Decline in Cardiorespiratory Fitness and Odds of Incident Depression. American journal of preventive medicine. 43(4): 361 – 368, 2012.
[27] Gorman, JM; Sloan, RP. Heart rate variability in depressive and anxiety disorders. American Heart Journal. 10(140): 77 – 83, 2000.
[28] Widjaja, D et al. Cardiorespiratory Dynamic Response to Mental Stress: A Multivariate Time-Frequency Analysis. Computational and Mathematical Methods in Medicine. Article ID 451857, 2013.
[29] Kralemann, B et al. In vivo cardiac phase response curve elucidates human respiratory heart rate variability. Nature communications. 4, Article ID: 2418, 2013.
[30] Möser, M et al. Phase – and Frequency Coordination of Cardiac and Respiratory Function. Biological Rythm Research. 26(1):100 – 111, 1995.

A principled approach to the optimization solution of the biometric system

Jernej Agrež
Faculty of Information Studies
Ulica talcev 3, 8000 Novo mesto, Slovenia
jernej.agrez@fis.unm.si

Miroslav Bača
Faculty of Organization and informatics
University of Zagreb
miroslav.baca@foi.hr

Nadja Damij
Faculty of Information Studies
Ulica talcev 3, 8000 Novo mesto, Slovenia
nadja.damij@fis.unm.si

Abstract: *The purpose of the research is to design a principled optimization solution for the loosely coupled and uncertain systems that would be applicable to the biometric system. We reviewed the state of the art to be able to design an ontology model of a biometric system. Further, we translated the model into relational databases that served us as a data source for the optimization tool. The solution operates as an open framework, tailored to the biometric system. It is capable of responding to its changes and developments and still provides a user with relevant support, either when operating with existing biometric system, or when designing a new multimodal biometric system.*

Key Words*:* *loosely coupled system, uncertainty, biometric system, optimization*

1 Introduction

Loosely coupled systems operate with high uncertainty, therefore optimization solutions applicable for such systems present an interesting research challenge. Even though the field of organizational optimization has already been studied in details, highly uncertain environments remain a topic that has not been studied to the full extent. In this paper, we present a principled approach to the biometric system, which can easily be included into the family of loosely coupled and uncertain systems as a specific case. On the first hand, many authors studied biometry and integrated different biometric methods into the system, that would cover the specific needs of recognition, control, etc. On the other hand biometric system`s optimization remains an open research topic. With the principled optimization solution, we try to reach following objectives. We try to develop approach, that would in the first place create the possibility to model the uncertain system in a transparent, intuitive and most importantly, open framework that is capable to capture such system, respond to its changes and supports the user with the optimization solution. We follow the objectives trough the three-tier optimization that consists of an open ontology framework, the transformation of the ontology into

relational databases, and finally, processing data with the optimization tool.

2 Loosely coupled and uncertain system optimization

Loosely coupled systems are conceptually widely used and diversely understood (Orton and Weick, 1990). Considering Glassman`s (1973) definition of such system, it is clear that it operates with high uncertainty, which is reflected through the reduced possibility to foretell detailed characteristic description of the task that we will attend to solve by the application of the system. Moilanen (2011) finds loose coupling and therefore the uncertain operational environment, as a stimulus for the organizational learning. Learning process as such creates a possibility to constantly upgrade system`s efficiency, with the support of the optimization tools. Li et al. (2013) argue that the crucial factor, when conducting system optimization, is to determine optimal process parameters. It is possible to achieve such determination through the learning process, that can gradually turn from human based learning to the automatized decision making. Liu et al. (2013) suggest a stepwise approach to the system optimization, starting with the definition of the problem and its characteristics. Further, we should consider what resources we would use to conduct a problem solving process and consider which, among appropriate resources are the most suitable for the practical application. To be able to sufficiently proceed with the selection, it is important to link all the resources into a graph and weight important links that represent relations among connected characteristics. Malakooti (2010) finds important also the ability to evaluate and rank different possible approaches to the problem solving solution, providing us with the insights, which, are among suitable system processes still acceptable, even though they do not demonstrate the highest suitability. With such measures, the optimization process becomes transparent and progressive, which is of high importance when we are dealing with loosely coupled systems, due to their uncertainty and the high possibility to become subjected to an unexpected change.

2.1 Biometric system

Benziane and Benyettou (2011) describe biometrics as a method for the identification of the individual person based on numerous characteristics. They find a biometric system as a highly reliable and fast identification solution. Dass et al. (2005) noticed application of the biometric systems in numerous different fields within the social environment, such as: travel, transportation, border control, homeland security, healthcare, banking and finance, access control, airport security, law enforcement, automotive, cyber security, encryption, nuclear power plants and watermarking. Elliot (2005) argues that different biometric approaches create the need for a different type of technology to such extend, that the technology roadmap is needed to be able to understand all the technological aspects and the possibilities within the field of biometry. Following the work of Schatten (2007), we can present the system as the list of methods on the one side, that consist out of statistical, physical, biological and other laboratory approaches. On the other side, we can find biometric characteristics, such as DNA, body odor, voice, signature, etc. Every included characteristic is defined also by the parameters such as acceptability, feasibility, permanence, measurability, etc. These parameters hold three qualitative values: low, middle and high. Most of the characteristics are submitted to the quality control, recognition and the structure extraction process. For each of these processes we use different methods, depending on the characteristic itself. Considering that, the assessed biometric system consists of 59 methods and 33 behavioral and physical characteristics that differ by 69 types of structure and 12 qualitative parameters. Not every biometric engineer has the ability to

follow technology roadmap, especially due to high costs. That is why it is of high importance to have the ability to optimize the system.

2.2 Biometric system optimization

To be able to approach to the biometric system optimization, we developed a process, built in three phases. In the first phase, we answered the need to visualize biometric system and put it into a transparent and easy to understand framework. Based on the theoretical insights we developed an OWL language based ontology, presented in Figure 1.

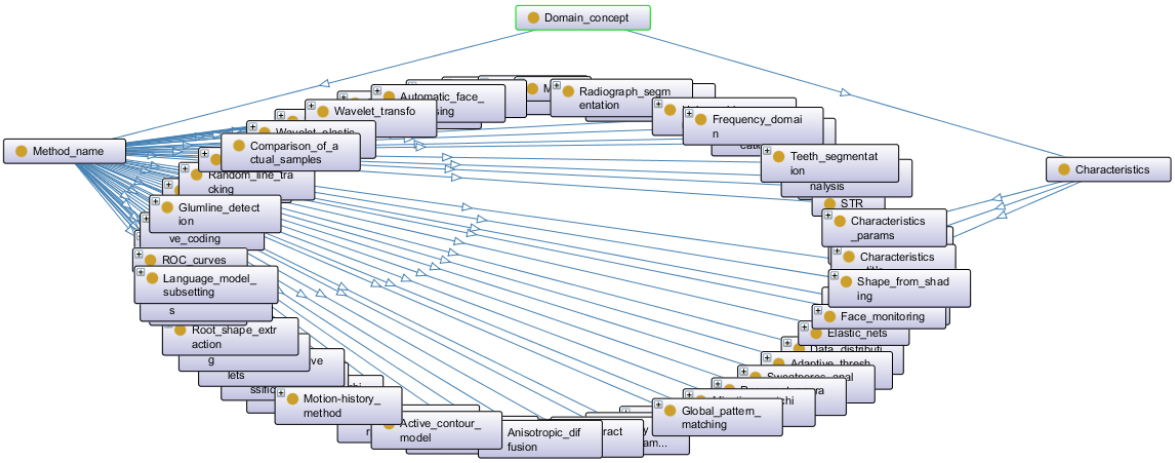


Figure 1: Ontology of the biometric system

We established the domain concept out of biometric characteristics and biometric methods. As presented in Figure 2, we further divided characteristics into characteristics` parameters, characteristic parameters` values, characteristics` structure, and characteristics` title. Second part of the domain concept were methods, presented in Figure 3.

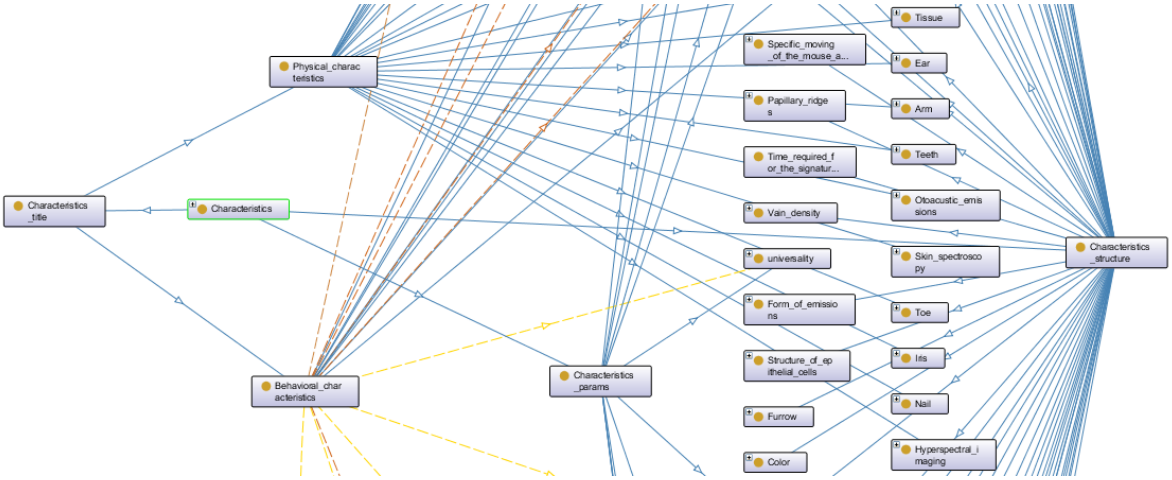


Figure 2: Characteristics in the biometric system

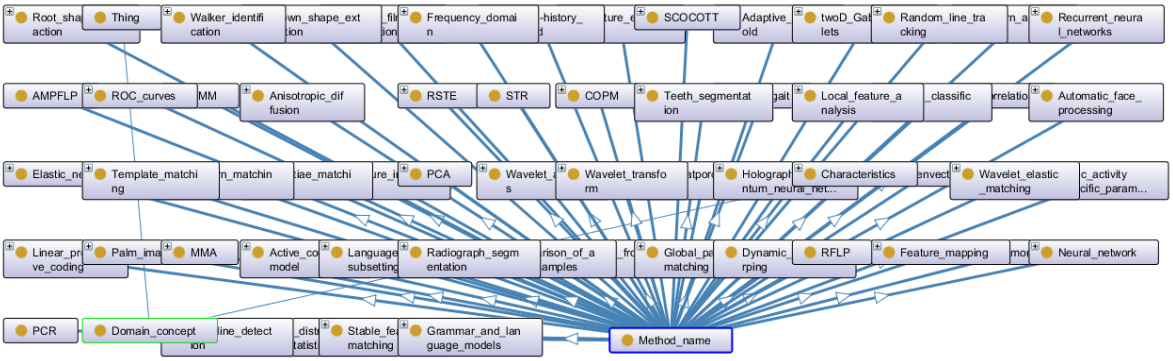


Figure 3: Methods in the biometric system

We selected characteristics, defined their parameters and connected them with methods, based on careful literature review. Having included all necessary elements of the biometric system, we needed to determine how to define links between the ontology elements and how to integrate relations, based on the links. We decided to use characteristics` title as a backbone of the linking process, because this element in most cases relates to all other elements within the ontology structure. Using ontology subclass as a frame for relating the elements, we connected every characteristic title with the type of structure, used for biometric purposes, with a feasible method for recognition, structure extraction and for quality control. We linked the characteristic with the behavioral or physical attribute and further linked it with the known assessment parameters. All relation links were established according to the theoretical insights from the literature review. To exceed bare linking we added to selected links also content weight that served as a relation indicator. When connecting characteristics with the available methods, in many cases, there is more than one feasible method and to reveal such case we linked characteristic to the method trough the relation “some”. If the characteristic used a single method for a specific purpose, we established a link with the relation “only”. Another relation that we needed to integrate into the link frame were the “high”, “middle” and “low” values that defined a set of characteristic parameters. We set the properties again as the subclass of the characteristic, using the same type of “some” and “only” relation to reveal which values correspond to which characteristic parameters.

In the second phase of the optimization process, we transformed ontology from its OWL form into five relation databases, where we included previously defined relations. The most comprehensive database presents relations between characteristics and their parameters together with their structure. The second database consists of the characteristic parameters, defining where high, medium, low values are labeled as positive, neutral or negative, and the last three databases contain relations between characteristics and methods for quality control, recognition and structure extraction.

Third and the most important phase in the optimization process was the design of a tool that would support user decision making, when facing the need to select the most appropriate methods out of the biometric system. According to the importance of designing the tool as an open framework, we choose to design it in an R programming environment. The tool is designed to proceed with the optimization in four steps. In the first step user selects the data source (relation databases) for the tool to work with. The importance of such procedure is the possibility of user to quickly, and simply add additional characteristics, methods and other elements to the data, if needed. In the following step user defines 10 most appropriate biometric characteristics that he

operates with, or that will take an important role in his biometric based solution. The third and the most comprehensive step consists of pondering the 10 chosen characteristics on the scale from 0.00 to 1.00. Pondering process is based on the following criteria. The tool firstly calculates how many positive, neutral, and negative parameters is held by a single characteristic. Further on it calculates the number of characteristic`s structures, and counts number of methods used for quality control, recognition and structure extraction. All the calculated values are then saved in the data frame, created by the programming tool. When the data frame is created (Table 1), weights are calculated for a single instance, adding 0.1 points to the weight, if characteristic has more positive parameters than neutral, if characteristic has more positive parameters than negative, and when it has more neutral parameters than negative. Further, in 0.1 points are added to the characteristic`s weight when its number of structures, structure extraction methods, quality control methods and recognition methods exceed mean of structure number within the data frame.

Table 1: Example of the data frame, used for optimization

positive	neutral	negative	structures	quality	recognition	extraction	weight	no
0	0	2	1	0	0	0	0.00	1
2	5	4	1	6	2	4	0.40	2
4	4	4	4	0	0	1	0.10	3
5	4	3	5	1	0	2	0.50	4
4	1	2	2	1	0	2	0.40	5
3	0	0	0	0	0	0	0.20	6
1	5	5	1	0	0	1	0.00	7
4	5	2	2	4	1	5	0.60	8
1	4	2	0	0	0	0	0.10	9
4	1	7	1	7	1	3	0.40	11

In the fourth step, weights were arranged on the scale from the lowest to the highest, revealing to the user, the value of the weight and serial number of corresponding characteristics.

3 Conclusion

This research includes insights about the optimization of the biometric system, which, due to its high uncertainty, holds characteristics of a loosely coupled organizational system. The uncertainty of the biometric system follows the state where systems, that are hard to predict to the fully extent, are being implemented. At the same time, it consists of several different criteria that define how to apply it in order to achieve most optimal results.

The main purpose of this research was to develop an optimization solution that would operate as an open framework and that could ensure selection of the most appropriate biometric approaches, according to the given situation. To be able to set up such solution, we decided to design it in two levels. At the first level, we created ontology, where we integrated all the elements of the biometric system and linked them according to the known rules and characteristics of the biometric system. Ontology provided us

with a transparent overview of the biometric system, allowing us to observe it from a holistic perspective, as well as to dive into the specific relations among different elements of the system. On the ground of the ontology, we translated the systems` elements and their relations in the several relation databases and further on developed optimization tool within the R programming environment. An optimization tool operates on the ground of the user`s selection of the biometric characteristic, and further on allocates them on the scale from 0.00 to 1.00. Allocation is assigned from the weights, calculated from the evaluated parameters of biometric characteristics, and quality control, recognition and structure extraction methods that are feasible with the selected characteristics.

Major conclusions of our work are being divided into two general aspects. A first aspect highlights an ontology as the primary stage of systems` optimization. Ontology provides us with the knowledge about the system and its elements. According to high uncertainty of the biometric system, it is necessary to model its structure within the ontology framework that is capable of linking its elements through the several different characteristics and other predefined connections, while being able to include any necessary change. A second aspect exposes the importance of the smooth downgrading conversion from the knowledge, retained in the ontology, into the data retained in the relation databases. It is necessary to proceed with such operation, due to the need to apply optimization, based on measurements of system characteristics. When we operate with the data, that describe the system in the details, optimization tool becomes a following framework that should be designed in an open manner as well. Openness is an attribute of high importance that provides us with the ability to include all necessary changes, originating on the first hand, from the change in the system, and on the other hand from the additional optimization needs.

For the future research, we recommend design of the solution that would be capable to conduct direct transformation from the OWL (XML) form ontology to the CSV form relation databases. At the same time, the existing optimization tool does not provide user with the ability to define his own pondering criteria, which would importantly increase usability of the optimization tool.

4 References

[1] Orton, J.D; Weick, K.E. Loosely coupled systems: A reconceptualization. The Academy of Managerial Review, 15(2): 203-223, 1990

[2] Glassman, R.B. Persistence and Loose Coupling in Living Systems. Behavioral Science, 18(2): 83-99, 1973

[3] Moilanen, S. Learning and the loosely coupled elements of control. Journal of Accounting & Organizational Change, 8(2): 136-159, 2011.

[4] Li, S; Liu, Y; Li, Y; Landers, R. G; Tang, L. Process planning optimization for parallel drilling of blind holes using a two phase genetic algorithm. Journal of Intelligent Manufacturing, 24(1):791-804, 2013.

[5] Liu; X; Yi, H; Ni, Z. Application of ant colony optimization algorithm in process planning optimization. Journal of Intelligent Manufacturing, 24(1): 1-13, 2013.

[6] Malakooti, B. Decision making process: typology, intelligence, and optimization. . Journal of Intelligent Manufacturing, 23(1): 733-746, 2012.

[7] Benziane, S; Benyettou. A. An introduction to Biometrics. International Journal of Computer Science and Information Security, 9(4), 2011.

[8] Dass, S. C; Nandakumar, K; Jain, A. K. A Principled Approach to Score Level Fusion in Multimodal Biometric Systems. Audio-and Video-Based Biometric Person Authentication, 1049-1058, Springer Berlin Heidelberg, 2005.

[9] Elliot, J. Biometrics roadmap for police applications. BT Technology Journal, 23(4): 37-44, 2005.
[10] Schatten, M. Zasnivanje otvorene ontologije odabranih segmenata biometrijske znanosti. Fakultet organizacije i informatike, Varaždin, 2007.

Automatic invoice capture in small and medium-sized Slovenian enterprises – final report

Darko Zelenika, Andrej Dobrovoljc, Robert Pezdirc, Helena Novosel, Simon Kegljevič, Janez Povh
Faculty of Information Studies
Laboratory of Data Technologies
Ulica talcev 3, SI-8000 Novo mesto, Slovenia
{darko.zelenika, andrej.dobrovoljc}@fis.unm.si, robert.pezdirc@elnino.si,
{novosel.helena, simon.kegljevich}@gmail.com, janez.povh@fis.unm.si

Bernard Ženko
Jožef Stefan Institute
Department of Knowledge Technologies
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
bernard.zenko@ijs.si

Božo Tomas
University of Mostar
Faculty of Mechanical Engineering and Computing
Matice hrvatske bb, 88000 Mostar, Bosnia and Herzegovina
bozo.tomas@hteronet.ba

Abstract: *In today's digital age paper documents are still in use and are a part of every enterprise. Many enterprises have a hard time managing all their paper documents, especially because this requires their manual transcription, which is time consuming. There are a variety of products on the market that offer solutions for automatic document capture and data extraction. Experience of Mikrografija d.o.o. suggest that small and medium-sized enterprises are slow in adopting these automatic solutions and the main reason for this is the high cost of such solutions. What most of them primarily want is to automate the capture process of their received invoices. The goal of Mikrografija d.o.o. is to provide such a service for an affordable price with a focus on small and medium-sized enterprises from Slovenia and other countries in the Adriatic region. They asked Faculty of information studies in Novo mesto to address these issues.*
In this paper we present the final results of this project and functionalities of developed invoice data capture software called mScan.

Key Words: *invoice data capture, OCR SDK analysis, analysis of the use of business applications*

1 Introduction

Even though nowadays computers and the internet are a part of practically every enterprise, and their documents mostly originate on the computer in a digital form, paper documents are still an integral component of their everyday tasks and are likely to remain so for the foreseeable future [5]. Invoice data capture is the most important task that enterprises usually perform on a daily basis mostly relying on human resources. However, manual data

capture is a very time consuming and expensive task and also a very error prone task. As the quantity of paper documents in different enterprises is growing daily, there is an increasing need to process these documents automatically with as little human interaction as possible. This automatic process saves a lot of time, money and reduces the required human resources.
The most popular solutions for automatic invoice capture on the market that enterprises can use are Abbyy FlexiCapture (Abbyy), Kofax Express (Kofax) and eFlow (TIS) [1, 7, 12]. These solutions use OCR (Optical Character Recognition) technology to transform paper documents to digital documents and then use the results obtained by OCR to perform data extraction.
The purpose of this paper is to introduce an invoice capture project which authors have worked on for 1 year and 3 months and report the final results. The main goal of the project overall is to offer to the Adriatic market an affordable and custom designed invoice capture product supporting languages from the Adriatic region. The analysis and evaluation of the available OCR SDKs [6, 14], analysis of Slovenian market regarding the use of business applications [10, 11] and introduction of new invoice data capture software (mScan) are the main contributions of this paper.
The rest of the paper is organized as follows. In section two, the invoice capture project is introduced. Section three describes the analysis of OCR SDKs. Section four describes summary of a survey on small and medium sized companies in Slovenia regarding the use of business applications. Section five describes development and functionalities of mScan invoice data capture software, while section six explains its architecture. Finally, section seven concludes the paper.

2 Invoice capture project

Mikrografija d.o.o. ("the contractor" in the rest of the paper) is a Slovenian company, which offers modern solutions for electronic document management and electronic archiving [8]. Their general activity is to capture, process and store paper documents in an electronic form.
The problem that the contractor faces is that small and medium-sized enterprises have a hard time in adopting automatic data capture software solutions which would convert their paper documents into electronic ones. The main obstacle is the price and complexity of such software. However, what most of these enterprises primarily want is to automate the capture process of their received invoices. Our contractor wants to offer a simple and easy to use invoice capture software solution tailored to the enterprises from the Adriatic region, which combines best functionalities of Kofax and Abbyy solutions into one product that is reasonably priced [15]. Research and development of the invoice capture software was offered as a project to the Laboratory of Data Technologies at the Faculty of Information Studies in Novo mesto, Slovenia [4]. Based on the contractor’s requirements the project is composed of the following phases - researchers from the Faculty of Information Studies need to:

- Analyze all the available open source and commercial OCR SDK solutions that enable development of the required software.
- Choose the best solution based on the software requirements.
- Investigate the Slovenian market regarding the use of business applications (ERP, CRM).
- Investigate the most frequently used metadata on the received invoices.
- Develop the application prototype.

In the following sections we describe each of these project phases in detail and report their results.

3 Analysis of OCR SDKs

An analysis of commercial and open source OCR SDKs [6] was conducted prior to the selection of Abbyy SDK as the main development tool. Based on specific needs of the contractor, multiple solutions were analyzed in search of the most appropriate OCR SDK. Ten commercial and two open source solutions were analyzed and evaluated [6, 14], which is shown in Table 1. Out of the twelve evaluated solutions, eight of them (1, 2, 3, 4, 5, 7, 8, 9 – Table 1) met the requirements of the contractor and three of the best solutions (1, 2, 3 – Table 1) were picked for further comparison. Abbyy OCR SDK had the highest rating of all analyzed solutions. Based on our research, open source OCR SDK solutions are comparable to commercial solutions. This is due to minimal implementation costs and satisfying functionality that open source solutions offer. Lower graded solutions did not provide the expected functionalities or were too expensive. Abbyy OCR SDK is one of most well-known solutions and offers a wide range of functionality at a high cost. Rated third, open source solution called Tesseract is free to use but lacks a few basic functionalities that other solutions offer. The most important missing features are picture preprocessing, ICR (Intelligent Character Recognition) and OMR (Optical Mark Recognition) support.
Besides already chosen Abbyy SDK the contractor decided to also include Tesseract (open source solution) into their product in order to offer multiple product packages.

Table 1: Results of OCR SDK analysis

#	Product	Points
1.	Abbyy OCR SDK	59.9
2.	Nicomsoft OCR SDK	45.3
3.	Tesseract (Ray Smith) - *open source*	42.4
4.	Dynamsoft Dynamic .NET Twain	42.5
5.	Puma.NET (Maxim Saplin) – *open source*	38.4
6.	LeadTools Recognition Imaging Developer Toolkit	38 (no SI[1] language support)
7.	Atalasoft DotImage	34.1
8.	Nuance Communications OmniPage	31.7
9.	Accusoft ImageGear	29
10.	Transym TOCR 4.0	20.6 (no SI[1] language support)
11.	ExpertVision OpenRTK 7.0	6.7 (no SI[1] language support)
12.	Aquaforest OCR SDK	4.5 (no SI[1] language support)

Source: own research

4 Summary of a survey on small and medium sized companies in Slovenia

Today, doing business is more and more electronic. Because of that companies need different program solutions than they have used in the past. ERP (Enterprise Resource

[1]Slovenian language support – criterion number two from Table 1.

Planning), CRM (Customer Relationship Management), BI (Business Intelligence) and similar, are the new age systems which are being used. The question we were trying to answer was if this is also the case in Slovenian companies. For foreign countries, we already know that working green, which means at least less paper business, is not just a trend but also a part of business policy. We wondered what the range of the electronic business in Slovenia is. How many documents are they saving in electronic versions (in DMSs (Document Management System) or in a cloud), and if and how they change paper documents to electronic form. For that purpose we performed a survey on small and medium sized companies in Slovenia. We received answers from 22 companies, although we sent the survey request to over 70 of them. In the survey, they were asked to answer on 15 questions [10, 11].

From the analysis of the answers we can conclude quite a few common traits of Slovenian business companies. For example, the most used business solutions are ERP and DMS, closely followed by CRM and BI. Mostly saved kind of documents are in PDF form, followed by Word documents and picture formats. Most of them are being saved in electronic databases. The largest amount of the surveyed use one or two databases. Two of the surveyed companies do not use a database. Those storages are on the companies premises and mostly connected with business solutions [10].

What we found out was that the Slovenian companies, as do the foreign, use more and more electronic data storage. They convert and create electronic documents, which enables them to access and use data from those documents faster and efficiently than ever before. Types of documents that are mostly electronized are from the finance group. And consequently, data that are most commonly used in later work are numbers. Which is logical, because profit and money is what makes the business and companies going. The data storage is mostly on the companies premises but there is a trend of cloud storage coming. To keep their data safe, they use various methods including user identification, which also allows the login entry data saving. For now, companies are very protective of their data and therefore do not want to risk the leakage of information with clouding. But as everything, trust in cloud storage will also change.

There are already signs of what could be improved to make doing business even easier, faster, safer and more economical. One of those things is a better connection of DMS and other business solutions which would make data access even more user friendly than it is today [11].

5 Development and functionalities of mScan's prototype

The mScan is software data capture solution which facilitates scanning, organizing and storing of documents. The basic use case scenario of the mScan is as following:

1. User imports documents from a scanner or file system to mScan
2. mScan automatically separates documents based on configured separation rules
3. User or mScan automatically (powerd by Abbyy FlexiCapture Engine) performs indexing (selection of document's metadata) of separated documents
4. User performs validation of indexed documents
5. User exports validated documents to the desired file format

The mScan is developed by using Microsoft .NET Framework 4.0 and C# programming language. The prototype of mScan is composed of cca. 14.000 lines of executable code.

In order to have modern graphical user interface (UI) and ribbon bar menu (Microsoft Office 2007 look) we used Telerik's WinForms UI controls [13]. In the following subsections we will describe each functionality of the mScan software.

5.1 Import of documents

Documents can be imported into mScan from two different sources: scanner and file system. To enable import of documents from scanner we used .NET Twain library [9] and we modified it for our purposes.

5.2 Separation of documents

Once documents are imported into mScan they can be automatically separated based on configured separation rules. The first rule is "As Origin File", which means if user imports multipage document from a file system into mScan it will be displayed as origin file i.e. one document multiple pages. The second rule is "Every N Pages", which means when a user imports multiple documents the application will create new document each time *N* number of pages are imported. The third rule is "Barcode", which means that the application will create new document each time a certain barcode type is recognized.
To enable recognition of barcodes from document images we used ZXing.Net [3] library, which enables recognition of 14 different barcode types.

5.3 Indexing and validation

Document indexing is a process of associating documents with different metadata (index fields) such as: invoice number, vendor name, date etc. Index fields enable us to search for documents more quickly.

5.3.1 Manual indexing

User can create many index fields by using form which is shown on the Figure 1. Each index field can be configured by using different parameters such as:

- Display name - Name of index field
- Default value - Default value for index field (username, recognized barcode value, etc.)
- On page - Number of page where index field is located
- Required - Index field is required (user is unable to export document if required field is not filled)
- Read only - Index field cannot be changed manually
- Validation - Validates index field value based on regular expressions
- Minimum length - Set desired minimum length of index field value to be valid
- Maximum length - Set desired maximum length of index field value to be valid

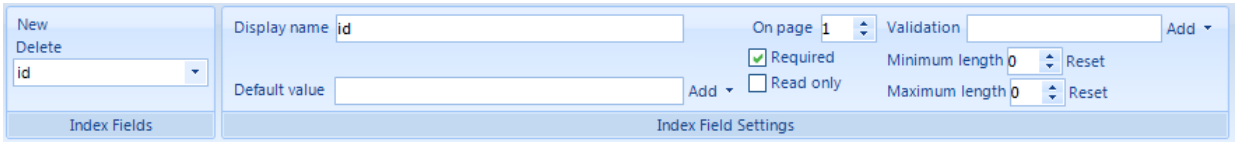


Figure 1: Index fields – creation form

After desired index fields are created and configured user is able to manually select index fields on a document image. Selected area is automatically processed by OCR Engine (Abbyy FineReader or Tesseract) and recognized text written into index field (Figure 2).

5.3.2 Automatic indexing

The mScan is powered by Abbyy's FlexiCapture Engine [2] which enables automatic indexing. FlexiCapture Engine uses Document Definition file to recognize and locate index

fields. Document Definition file describes: how to identify a document, what data need to be recognized and how to find this data [2, 14]. Therefore, Document Definition represents a simple document classifier, which has to have unique features, which are used to recognize documents. Document Definition file can be created by Abbyy's tool which is called FlexiLayout Studio and it can be used in mScan. Once Document Definition file is imported into mScan all fields from Document Definition file are transformed into mScan's index fields and during indexing process are automatically located and selected on document images (Figure 2).

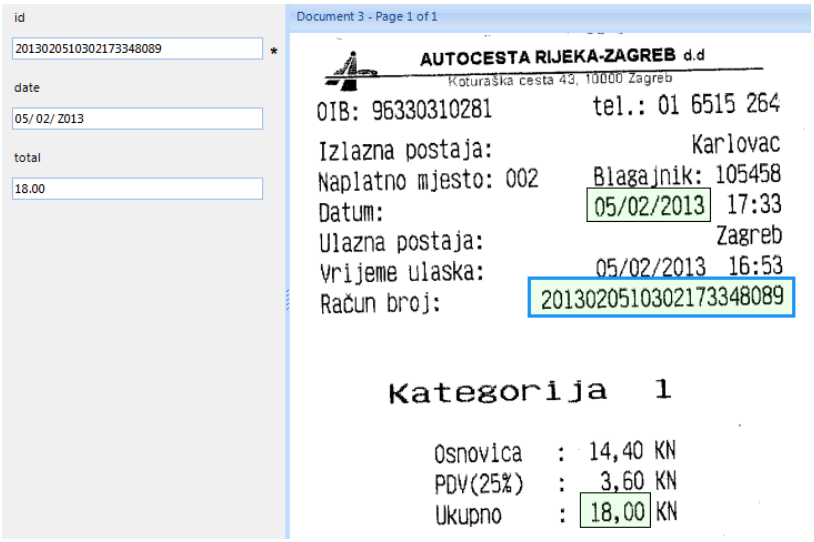


Figure 2: Manual indexing

5.3.3 Validation

After indexing of documents user is able to validate recognized index fields manually by the help of validation rules. If index field value does not fit configured validation rules user will be warn to correct index field value.

5.4 Export of documents

After validation process user is able to export documents. The mScan enables export of documents to file system or to the contractor's cloud system – mSafe [8]. Documents and index fields can be exported separately. Documents can be exported to TIFF and PDF file formats. To enable export to TIFF file format we used N. Fishman's TifUtils library and for PDF export we used Abbyy's FineReader Engine. Index fields can be exported to XML, CVS and TXT file formats.

6 Architecture of mScan

The mScan is build based on plug-in architecture. Plug-in architecture allows developers to build applications that are easily extensible. Figure 3 depicts mScan's architecture. We created the host application in which we defined rules for every feature of the mScan that we want it to be extensible. Therefore, we defined rules for:

- Import plug-ins - we developed three plug-ins that enable import of documents
- Document separation plug-ins – we developed one plug-in that enable separation of documents
- OCR plug-ins – we developed two OCR plug-ins (commercial and open source)

- Document indexing plug-ins – we developed one plug-in that enable automatic document indexing
- Export plug-ins – we developed two plug-ins that enable export of documents
- Export file format plug-ins – we developed two plug-ins that enable export of documents to various file formats. These rules are rules for export plug-ins, which means that each export plugin will support export to these file formats.

Application developed in such way enable adding of new features to the mScan without accessing source code.

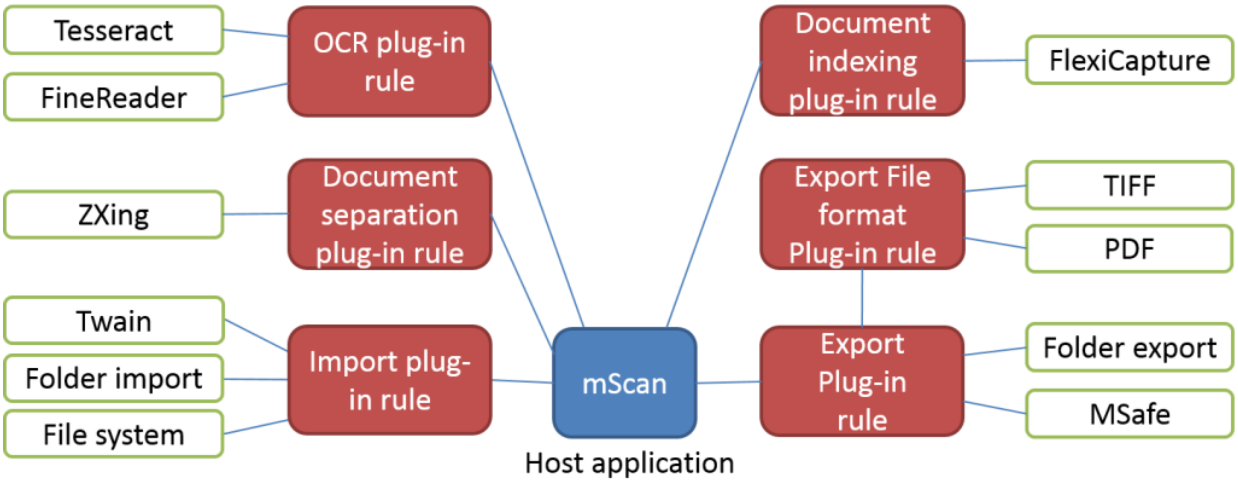


Figure 3: mScan's architecture

7 Conclusion

Automatic data capture software can be very helpful and time-saving for enterprises. Small and medium-sized enterprises are mostly dealing with invoices and often refuse the use of such software because of its complexity and price. Authors of this paper were part of the project whose aim was to develop affordable invoice capture software for small and medium-sized enterprises from the Adriatic region. Based on the analysis of twelve open source and commercial OCR SDKs, the highest rating was given to Abbyy SDK, which was also chosen as the main development tool for our project, together with Tesseract best rated open source solution. From the analysis of Slovenian market we found out that the Slovenian companies, as do the foreign, use more and more electronic data storage. Types of documents that are mostly electronized are from the finance group. And consequently, data that are most commonly used in later work are numbers. We introduced and showed all functionalities of the invoice data capture software mScan which will soon be available on the market.

8 Acknowledgements

The presented work was supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

9 References

[1] Abbyy - FlexiCapture, http://www.abbyy.com/data_capture_software/, downloaded: October 2nd 2013.

[2] Abbyy - FlexiCapture Engine, http://www.abbyy.com/flexicapture_engine/, downloaded: October 2nd 2013.

[3] Codeplex, https://zxingnet.codeplex.com/, downloaded: August 14th 2014.

[4] DataLab,Laboratory of Data Technologies – Projects, http://datalab.fis.unm.si/joomla/index.php/projects, downloaded: October 2nd 2013.

[5] Jervis, M; Masoodian, M. Evaluation of an Integrated Paper and Digital Document Management System. In Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction – INTERACT 2011, pages 100-116,Lisbon, Portugal, 2011.

[6] Kegljevič, S. Analysis and comparison of existing OCR libraries. Bachelor's thesis, Faculty of Information Studies, Novo mesto, Slovenia, 2013.

[7] Kofax, http://www.kofax.com/document-scanning-software, downloaded: August 14th 2014.

[8] Mikrografija d.o.o, http://www.mikrografija.si/, downloaded: October 2nd 2013.

[9] NETMaster. Code Project, http://www.codeproject.com/Articles/1376/NET-TWAIN-image-scanner, downloaded: August 14th 2014.

[10] Novosel H. Needs and expectations of small and middle sized companies with the capture of document content. Bachelor's thesis, Faculty of Information Studies, Novo mesto, Slovenia, 2014.

[11] Pezdirc R. The prevalence of business information solutions and the need to connect to documentation systems. Bachelor's thesis, Faculty of Information Studies, Novo mesto, Slovenia, 2014.

[12] The eFlow platform, http://www.topimagesystems.com/solutions/overview/eflow-overview, downloaded: October 2nd 2013.

[13] Telerik. Telerik, http://www.telerik.com/products/winforms.aspx, downloaded: August 14th 2014.

[14] Zelenika, D.; Kegljevič, S.; Dobrovoljc, A.; Povh, J.; Ženko, B.; Tomas, B. Automatic invoice capture in small and medium-sized Slovenian enterprises – project overview. In Proceedings of the 5th International Conference on Information Technologies and Information Society (ITIS 2013), pages 40-46, Dolenjske Toplice, Slovenia, 2013.

[15] Zelenika, D; Povh, J; Dobrovoljc, A. Document Categorization Based On OCR Technology: An Overview. In Proceedings of the 7th European Computing Conference, pages 409-414, Dubrovnik, Croatia, 2013.

Project of manufacturing processes optimisation in Podgorje Ltd.

Tadej Kanduč, Blaž Rodič
Faculty of Information Studies
Sevno 13, 8000 Novo mesto, Slovenia
{tadej.kanduc, blaz.rodic}@fis.unm.si

Abstract: *In this paper we present a project of optimising the manufacturing processes in Podgorje Ltd., a Slovenian furniture company. The key steps of the project, such as extracting and preparing data, synchronising different software environments, building simulation model and optimisation process are outlined in this paper.*
We have constructed a discrete event simulation (DES) model of manufacturing processes, which allows us to understand the current process and optimise its parameters. We have also developed a method for automated model construction which uses the listed data to modify the simulation model. At the end we present the optimisation problem handled in the project: minimisation of the total distance the products need to travel between the machines, which we solved by optimising the machine layout on the factory floor using an automated heuristic method.

Key Words*: discrete event simulation, optimal layout, heuristic optimisation*

1 Introduction

Manufacturing processes in larger production companies are generally complex and need to be systematically organised in order to achieve high levels of efficiency. Companies need to consider several criteria and restrictions in the processes such as costs, due dates, amounts of stock materials, different measurements in efficiency, etc.

Methods suited for modelling of complex manufacturing systems usually include discrete event simulation (DES) modelling. Construction of a DES simulation model requires that the data that describe the manufacturing processes are obtained, analysed, extracted and prepared in a suitable format for the model.

System or process optimisation can be performed by implementing changes in the model, usually by constructing several versions of the model and input data (i.e. scenarios) and comparing simulation results.

In the paper we present main steps of the project of optimising manufacturing processes in Podgorje Ltd., a Slovenian furniture company. Primary goal of the company is to reduce overall costs in manufacturing processes. Our goal was to investigate how the layout of machines on the factory floor affects the efficiency of manufacturing processes. Our primary optimisation criterion was the total distance the manufactured products need to travel on the floor, however we have also monitored various other criteria during the optimisation processes. The results of our project are used within a currently running micrologistics optimisation process.

1.1 Problem situation

Approximately 140 machines are located on the factory floor of the company. The company catalogue contains more than 30,000 different products. Each product is manufactured according to the prescribed bill of materials (BOM) and its technical procedure. In BOM, required semi-finished products and materials to manufacture the given product are listed. The technical procedure data describe the sequence of operations in the manufacturing of the given product. Manufacturing processes include large set of different products and variations of open orders during each working month. Production scheduling is based on customer orders and performed using the Preactor scheduling system.

1.2 Previous research (review of literature)

Simulation is commonly used for the evaluation of scenarios [1],[2],[3]. However, the models developed with the visual interactive modelling method (VIM) are usually manually constructed through careful analysis of the real-life system and communication with process owners. Automated model development is more common with methods that allow easier and more standardised formal description of models, e.g. Petri nets [4],[5]. Automation of model construction and adaptation can importantly facilitate the development of models of complex systems [6],[7] and generation of simulation scenarios.

Several papers deal with factory layout optimisation, with paper [8] stating that multiproduct enterprises requires a new generation of factory layouts that are flexible, modular, and easy to reconfigure. Evolutionary optimisation methods are often proposed due to problem complexity [9]. Layout optimisation problem is identified as hard Combinatorial Optimization Problem and the Simulated Annealing (SA) meta-heuristic resolution approach is proposed to solve to problem [10]. A novel particle swarm optimization method is proposed by [11] for intelligent design of an unconstrained layout in flexible manufacturing systems.

Factory layout design optimisation is further discussed in [12],[13],[14]. Authors [12] propose a new facility layout design model to optimise material handling costs. Sources [13] and [14] propose genetic algorithm based solutions to respond to the changes in product design, mix and volume in a continuously evolving work environment.

2 Methodology

Our task in the project was to develop a better machine layout, which will fit the current production needs and projections for the next ten years. To complete this task, we have developed a simulation model of the factory floor and optimisation methods based on the company data and their specific optimisation goals. The purpose of the model is verification of new manually or algorithmically generated floor layouts. For most of the machines there are no specific location restrictions. It is also be possible to add some new machines on the floor if considerable improvements can be achieved.

Optimisation of floor layout is conducted in cooperation with experienced manufacturing planners, managers and other experts within the company, and is facilitated by state-of-the-art optimisation algorithms that are employed to generate new layout scenarios, i.e. to search for the optimal layout within a large set of possible layouts.

2.1 Existing tools and data

As a part of established scheduling and planning procedure, Podgorje Ltd. uses Preactor software (http://www.preactor.com/) to schedule customer orders according to a set of priorities and availability of resources (machines) and daily monitor manufacturing processes on the factory floor. Preactor is a family of "advanced scheduling and planning" products that allows detailed definition of manufacturing and other processes, and integrates with existing ERP and other company databases and applications.

However, the modelling process within Preactor is not flexible enough to allow easy modification of the system model or modelled processes and testing of scenarios, required for layout or process optimisation. To simulate processes in a different factory floor layout, an entire simulation model needs to be built from scratch or undergo lengthy manual modification.

2.2 Selection of tools and methods

We decided to implement current production processes and optimisation procedure with a specialized simulation and modelling tool Anylogic – a powerful software that implement DES, SD and agent based modelling (ABM) methodologies. Modelling is performed using VIM approach which is intuitive and clear, and it supports advanced visualisations techniques. Anylogic or other simulation and modelling tools are not a replacement for advanced scheduling and planning tools as Preactor or vice versa. Instead, they complement each other: Preactor contains a detailed process model that allows accurate scheduling and planning and provides detailed process data for Anylogic, while Anylogic allows fast design and optimisation of processes, addition of new machines and verification of scenarios using different factory layouts and sets of orders.

The Anylogic simulation model allows us to monitor various manufacturing process statistics and to better understand the manufacturing system by discovering rules and connections in the manufacturing system. The model was verified by comparing the simulation results (e.g. manufacturing time, machine utilisation) using synthetic and real historic order data prepared by the company planners with the real-event statistics of the set of orders from the past year.

An important part of the project was the preparation and export of manufacturing process data and customer order data from the company database, and the connection of all software components (databases, simulation model, model construction application and auxiliary applications) in order to achieve the required level of integration.

2.3 Data based modelling

Manufacturing process data includes the data for technical procedures and BOM. They are stored in Microsoft SQL Server database, used mainly by the manufacturing scheduling application Preactor. Preactor is used in Podgorje for online monitoring of the manufacturing processes via approximately 100 control points and to schedule manufacturing orders based on customer orders. This allows the company to daily update and, if necessary, modify the manufacturing schedule.

We have analysed the structure and content of database tables and prepared a set of queries that were used to extract the data required for model construction and simulation scenario generation, i.e. the preparation of model input data. The queries were stored in the Microsoft SQL Server database in the form of views and later called by an Excel workbook. The workbook was used as an intermediate data storage that allowed us to examine and modify the data as required. Some corrections were necessary as the original database contained some errors and some data was missing for certain technical

procedures and BOMs. This is an inevitable step when dealing with real-life data. Table 1 shows an example of an SQL query used to obtain the data on machines in machine groups (referred to as Resources and ResourceGroups).

Table 1: Example of an SQL query

```
/*Podgorje_baza_20140403.LSI.*/
CREATE VIEW Test19Projects_equivMachines AS
        SELECT ResourceGroupId, RGR.ResourceId, ResourceCode FROM
                Podgorje_baza_20140403.LSI.ResourceGroupResources RGR,
                Podgorje_baza_20140403.LSI.Resources R
        WHERE RGR.ResourceId=R.ResourceId
;
```

Anylogic stores the models as standard XML files, which allows easy manual or algorithmic modifications of the model. To this end we have developed an application in Java that reads input data from Excel and constructs the corresponding Anylogic model by modifying a template model.

2.4 Input and output data

All the input data (orders, technical procedures, BOMs, etc.) are primarily stored in SQL databases, generated by Preactor software. Relevant data are saved as queries and exported to intermediate Excel file. In Excel, the data are slightly manually modified, since inaccurate and inconsistent in real data occasionally occur. In Excel, the following input data are stored:

- An order is described as a list of products (catalogue numbers). For every product from the list, name, quantity, earliest start time, priority parameter and volume are assigned.
- Each product has a specific technical procedure. For every operation there is a group of equivalent machines, a preferred machine, set up time and time per item.
- More complex products also have bill of materials, i.e., list of required semi-finished products or materials that are joined at a specific operation in specific quantity.

At start-up of the simulation, input data from Excel are read and stored in internal Anylogic arrays. From there on, all data are read from internal data structures to remove constant communication with external files, which would slow down the simulation.

During simulation, various statistical data are measured and stored:

- For every pair of machines, different types of flows (number of products, number of used carts, overall volume of products and total distance of carts) are measured.
- For every machine, utilisation, overall setup time, flow of products and volume, and queue of products are monitored.
- For each series of products, completion times and sequences of machines, which were chosen during simulation, are stored.
- Different, less significant measurements, such as flow of carts and routes of the carts, are recorded.

Once the simulation is finished, all the data are stored in the output Excel file,

2.5 Components of the simulation system

Modelling and simulation system is composed of four main elements:

- Core manufacturing process simulation model in Anylogic environment.
- Java application that constructs XML Anylogic model from a template file.
- MS Excel as an intermediate input and output data storage, and analysis tool.
- MS SQL server database describing technical procedures and client's orders.

The resulting system is shown in Figure 1. The simulation run is prepared as follows. First, we prepare Anylogic template file (XML). Simulation model (new Anylogic XML file) is constructed by running the Java algorithm for automatic model building. Next, we run the Anylogic simulation model. During simulation, input database is read dynamically. When simulation is finished, simulation results are stored in output Excel file.

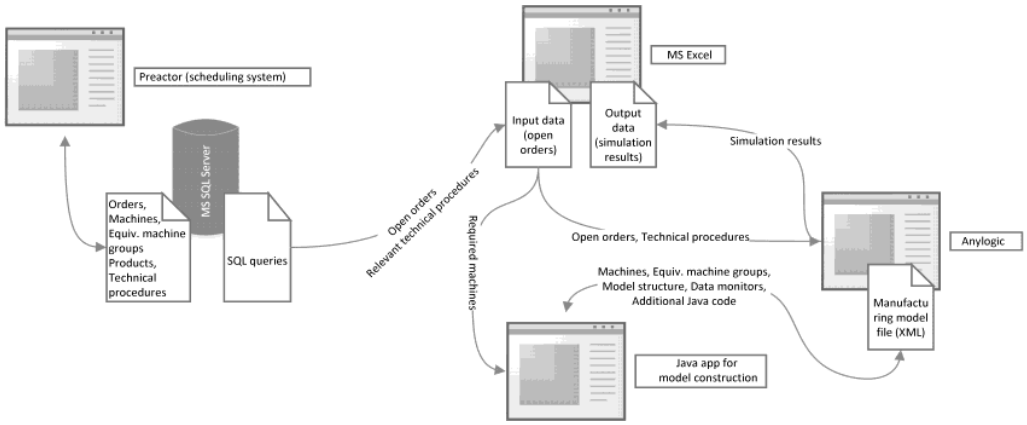


Figure 1: System schematics

2.6 Optimisation methods

In this section we describe the problem of finding the factory floor that minimises total transportation distances of the products during the production. We have tried different optimisation algorithms to minimise the total distance of the products. We have tested freely available open source heuristic algorithms in C++ and Matlab for quadratic assignment problem that are based on simulated annealing [15], iterative local approach [16] and ant colony algorithm [17]. We have also designed a heuristic optimisation method using the system dynamics (SD) methodology in Anylogic that is based on force-directed drawing algorithms. So far, it has generated good results and will be further developed and explored.

The optimisation problem is presented as finding the optimal mathematical network, in which nodes of the network represent the machines on the factory floor and weighted edges between the nodes represent transactions between the machines. Real routes on the floor between the machines are neglected in this case, since it considerably complicates the optimisation problem. The optimisation method should only propose a basic outline of the layout, since the final layout needs to be further tuned by the company experts to meet other less precise criteria.

Each machine takes specific amount of space on the floor and machine regions must not intersect. If we further presume that all machines take the same amount of space

on the floor, we can restrict the machine positions to discrete points on a predefined grid. Hence the problem simplifies to well-known quadratic assignment problem (QAP).

As an alternative to QAP algorithms, we have developed a promising alternative optimisation method, which is based on force-directed graph drawing methods. To every machine m_i we prescribe the corresponding repelling force F_{ij} to every other machine m_j. Repelling forces keep the machines away from each other since we want sufficient space between the machines. For every pair of machines m_i, m_j we define an attractive force G_{ij}, which is proportional to the weight f_{ij} and the distance $dist(m_i, m_j)$. Attractive forces move the machines with larger volume transactions closer to each other. The machines are repositioned according to the defined forces in the system. When the machines do not move any more, the system has reached a local minima.

3 Results and discussion

The main outcomes of the project are the integrated simulation model in Anylogic that communicates with external database files, the method for automatic model construction and the novel heuristic floor layout optimisation method. The model servers as an indispensable tool for in-depth analysis of the manufacturing process.

The novel optimisation method outperformed other more general heuristic methods for QAP in terms of the optimisation criterion. The generated layout has approximately 30% shorter total product travel distance than the current layout. Shorter travel also means less time the workers need to transport of products. The customer has responded very favourably to these results, and prepared several manually adjusted floor layout based on our generated layout that will be verified with the simulation model before selection and implementation.

Further steps in our project will include changing the set of machines: replacement of one or several machines by newer multipurpose CNC machines. Other optimisation goals and criteria will be explored.

4 Acknowledgements

Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

5 References

[1] M. Kljajić, I. Bernik, A. Škraba: Simulation approach to decision assessment in enterprises. Simulation, 2000, 75(4), pp. 199-210.

[2] R. S. Edis, B. Kahraman, O. U. Araz and M. K. Özfirat: A facility layout problem in a marble factory via simulation. Mathematical and Computational Applications, 2011, 16(1), pp. 97-104.

[3] P. Tearwattanarattikal, S. Namphacharoen and C. Chamrasporn: Using ProModel as a simulation tools to assist plant layout design and planning: Case study plastic packaging factory. Songklanakarin Journal of Science and Technology, 2008, 30(1), pp. 117-123.

[4] R. Conner: Automated Petri net modeling of military operations, in IEEE Proceedings of the IEEE 1990 National Aerospace and Electronics Conference - NAECON 1990, Volume 2, 1990, Dayton, Ohio, USA, pp. 624-627.

[5] D. Gradišar and G. Mušič: Automated Petri-Net Modelling for Batch Production Scheduling, in (ed: Pawel Pawlewski) Petri Nets - Manufacturing and Computer Science, InTech, 2012, pp. 3-26. http://dx.doi.org/10.5772/48467

[6] A. D. Lattner, T. Bogon, Y. Lorion and I. J. Timm: A knowledge-based approach to automated simulation model adaptation. In Proceedings of the 2010 Spring Simulation Multiconference (SpringSim '10). Society for Computer Simulation International, San Diego, CA, USA, 2010, Article 153.

[7] R. Kannan and H. Santhi: Automated construction layout and simulation of concrete formwork systems using building information modeling, In (eds: Djwantoro Hardjito & Antoni) Proceedings of The 4th International Conference of Euro Asia Civil Engineering Forum 2013 (EACEF 2013), National University of Singapore, 26-27 June 2013, pp C7-C12.

[8] S. Benjaafar, S. S. Heragu and S. A. Irani: Next generation factory layouts: Research challenges and recent progress. Interfaces, 2002, 32(6), pp. 58-76.

[9] A. Sadrzadeh: A genetic algorithm with the heuristic procedure to solve the multi-line layout problem. Computers and Industrial Engineering, 2012, 62(4), pp. 1055-1064.

[10] G. Moslemipour and T. S. Lee: Intelligent design of a dynamic machine layout in uncertain environment of flexible manufacturing systems. Journal of Intelligent Manufacturing, 2012, 23(5), pp. 1849-1860.

[11] M. Ficko, S. Brezovnik, S. Klancnik, J. Balic, M. Brezocnik, and I. Pahole: Intelligent design of an unconstrained layout for a flexible manufacturing system. Neurocomputing, 2010, 73(4-6), pp. 639-647.

[12] K. K. Krishnan, S. H. Cheraghi and C. N. Nayak: Facility layout design for multiple production scenarios in a dynamic environment. International Journal of Industrial and Systems Engineering, 2008, 3(2), pp. 105-133.

[13] J. S. Kochhar and S. S. Heragu: Facility layout design in a changing environment. International Journal of Production Research, 1999, 37(11), pp. 2429-2446.

[14] M. Enea, G, Galante and E. Panascia: The facility layout problem approached using a fuzzy model and a genetic search. Journal of Intelligent Manufacturing, 2005, 16(3), pp. 303-316.

[15] E. Taillard: Simulated annealing (SA) procedure for the quadratic assignment problem, 1998, http://mistic.heig-vd.ch/taillard/codes.dir/sa_qap.cpp.

[16] S. Shah: Implementation of iterative local search (ILS) for the quadratic assignment problem, http://shah.freeshell.org/ilsassignment/ cILSAssignment.cpp.

[17] K. Tsourapas: Ant Algorithm for the Quadratic Assignment Problem, 2008, http://www.mathworks.com/matlabcentral/fileexchange/1663-qap.

Modern IT solutions in the logistics process

Daniel K. Rudolf
Actual I.T. d.d.
daniel.kovacecicrudolf@actual-it.si

Abstract: *In the IT systems development in the logistics port area operations, the company Actual IT is faced with the shipper agents need to have an insight into the status of their transmitted messages (eg. working orders) in the Luka Koper system. Therefore, the company wants to determine which technologies are most appropriate for transmission and transmitted message viewing. The aim of this talk is to determine which tools for message intervening and reviewing are best suited for displaying the computer messages. The results showed that for the making electronic message viewer is preferred solution in the website form built with the MVC architectural pattern and using Solr search technology.*

Key Words*: Solr, EDI, ASP.NET, Message Broker, Message Viewer, Logistic process…*

Information security culture of online banking users

Andrej Dobrovoljc, Tomaž Perko, Jože Bučar
Faculty of Information Studies
University of Novo mesto
Sevno 13, 8000 Novo mesto, Slovenia
{andrej.dobrovoljc, joze.bucar}@fis.unm.si

Abstract: *Information technology improves the quality of our lives on various fields. Online banking is one of the most widely used services today and is one of those things, which has drastically influenced our daily habits. We can make banking transactions from home or some other location at any time and from different computer devices. Unfortunately, together with comfort, such solutions brought us also some serious risk. Intruders have uncovered numerous weaknesses of web-based services and consequently they find them attractive for their malicious plans. Many studies prove that people are the weakest part in the chain of security. It is not enough for people only to understand the security threats around these modern solutions but they have to behave appropriately as well. We are interested in information security culture of online banking users. Therefore, we developed the measuring model and applied it on online banking users in Dolenjska and Bela Krajina.*

Key Words*:* *security culture, online banking, risk*

1 Introduction

Web banking services compared to the classical ones offer many benefits. They outperform the classic approach in accessibility, comfort and costs. Therefore, the vast majority of banks all over the world today offer the online banking to their clients. Consequently, people have bigger opportunity to choose the most suitable bank because they are globally present and accessible from anywhere. We can conclude that both sides, clients and banks, should profit out of it.

The first estimates that the web banking would cut the bank business costs up to 25% was an exaggeration. In reality, it is somewhere at about 5%. Today, the web bank is not an advantage any more but a must, otherwise it is not attractive. It seems that clients profited more, because they can access bank 24 hours a day from any location and by using various mobile and computer devices.

In Slovenia, web banking for people started around the year 2000. In 2002, there were 98.669 registered users. Up to the year 2006 this number jumped to 351.111 users what is 256% more than in 2002. The main reason for the steep growth was better availability and higher speeds of internet. During these years, users accepted the web banking solution and found it as one of the key services on the internet. In 2013 the number of users was 690.040, what is 96,5 % more that in 2006. These data prove that the online banking market is approaching saturation. The web based banking is widely accepted today.

Unfortunately, web banking has also some serious drawbacks. Classical banking approach was safe, because we made all transaction at the bank counter and ordered transactions to the bank clerk. With the online banking we should be careful whether we are on the genuine bank web site and whether we got the messages from the bank and not from the intruder. Banks are aware of such security risks. Therefore, they introduce various security measures and inform their clients about importance of secure behaviour. The question is how clients perceive security risks and how do they obey the rules of secure behaviour. It is very dangerous, if someone understands the security risk and importance of security measures, but he or she does not behave accordingly. In this situation, we can speak about the information security culture. We define it as the difference between the information security knowledge and the actual behaviour of users.

In case of online banking, a high information security culture is of utmost importance, because people can suffer a serious harm. Our intention was to measure the information security culture of users of online banking in Dolenjska and Bela Krajina. We developed a measuring model, which is supported by the questionnaire. Our key finding is that the current security culture does not reach the sufficient level.

2 Related work

The information security culture has developed from security culture. It covers the specific environment of informatics, where just the application of various technological security measures is not enough (Rančigaj & Lobnikar, 2012). Security assurance of IS in a great deal depends on behaviour and participation of people (Mitnick & Simon, 2002). Perception of importance of security, knowledge about security measures and ethics to act according to this knowledge are the key human factors, which support the high culture (Rančigaj & Lobnikar, 2012). Without the necessary knowledge and the suitable behaviour, we cannot reach the desired security despite the best possible technological security solutions.

Some researchers (Da Veiga & Eloff, 2010; Ella Kolkowska, 2011; Niekerk & Solms, 2006; Ruighaver, Maynard, & Chang, 2007) defined their information security culture models above Schein's definition of organizational culture. It consists of three levels:

- **Artefacts** (visible part): behaviour patterns, technology, forms.
- **Values** (partly visible): official documents, which describe values, principals, ethics, and visions.
- **Assumptions** (hidden): people's convictions about how the system works.

Da Veiga and Eloff (Da Veiga & Eloff, 2010) define the information security culture as a set of values, behaviour, assumptions, beliefs and knowledge of all the information system (IS) stakeholders. Their interactions with the IS can be acceptable or inacceptable and they result in a specific form of security assurance, which varies in time. Van Niekerk & Von Solms added the fourth level, to the Schein's model and named it "information security knowledge" (Okere, Niekerk, & Carroll, 2012). This level supports the first three levels and thus ensures compliance.

3 Defining the model

Existing studies are focused on measuring the information security culture within the organization. In case of online banking, we speak about the culture within the certain group of users, who may not have a direct relationship as it is within the organization, but they definitely share some experiences of usage in various informal or formal ways.

A responsible user of online banking needs to develop his attitude toward the usage. In order to achieve sufficiently high level of culture, he or she has to acquire the knowledge at all levels of information security culture. Therefore, we extended the Niekerk & Von Solms four level model so that the knowledge level is divided into three parts. Each one is related to the corresponding levels from the original Schein's model. Our intention is to measure the knowledge about the individual level at first, in the second step to measure the behaviour and in the end to calculate the difference. By our definition, the difference between the level of user behaviour and the level of his or her knowledge represents the security gap. When the knowledge is good and the security gap is small, the information security culture is high and vice versa.

The key difference between the knowledge and the behaviour is the action. People with the same level of knowledge can react differently. Therefore, we have to measure both constructs separately. Questions, which are related to the constructs at the same culture level, are similar (Table 1). The only difference is the focus of the questions. Questions regarding the knowledge reveal us the current level of user competences. On the other hand, the questions regarding the behaviour reveal how seriously users take the knowledge about security.

Table 1 Information security culture model for measuring the online banking service

Information Security culture levels	User knowledge and perception of security	User behaviour, values and assumptions
Artefacts	How should we protect?	How do I protect?
Values	What should we protect?	What have I protected?
Assumptions	Why should we protect?	Why do I protect / not protect?

However, user knowledge about security and its proper perception are the key factors to achieve a high information security culture. The service provider has to invest into awareness and education of clients about general principals of safe usage.

4 Method

The main purpose of this study was to measure the information security culture of online banking users in Dolenjska and Bela Krajina. We define that the information security culture of online banking is high when on average the knowledge about security and people behaviour are above 80% of maximum possible value. This definition follows the Pareto principal, which says that we can achieve the result of 80% with an effort of 20 %. The question is if the society has invested enough efforts for a high information security culture of online banking users. In this sense, we place the following hypothesis:

> H1: The information security culture in Dolenjska and Bela Krajina does not reach sufficiently high level, which is 80% of maximum possible value (the average of all three culture levels together).

H2: The information security culture is better in Dolenjska than in Bela Krajina.

H3: The knowledge about security is on average higher than people behaviour.

The measurement instrument is questionnaire based. There is one question for each construct from the proposed model and a 5-step Likert scale is used. At the end of the questionnaire some demographic questions are added. We carried out an online survey. The sample was selected by a snowball method.

5 Results and discusion

In the survey participated 260 persons, but only 138 questionnaires were completely and correctly filled out. Partly completed questionnaires were removed from the statistical analyses. Among participants, there were 63% male and 37% female. Some participants refused to answer the demographic question related to education level. Table 2 presents the education level of participants in the sample and Fig. 1 the age structure.

Table 2 Education level of participants

Education	**Share**
Less than 4 year middle school	5 %
Middle school (4 years)	42 %
Upper school (2 years)	15 %
Bachelor's degree (1. Bologna level)	10 %
Bachelor's degree (before Bologna), Master degree (2. Bologna level)	21 %
Master degree (before Bologna)	3 %
Doctoral degree	4 %

The education structure match with the statistical data for Slovenia in 2013. Fig. 1 proves that online banking is not just a domain of younger people. The majority of users belong to the age group 41-50 year. However, elder people (51-60 year) started using online banking as well, despite the fact that IT is something new for them and demands learning.

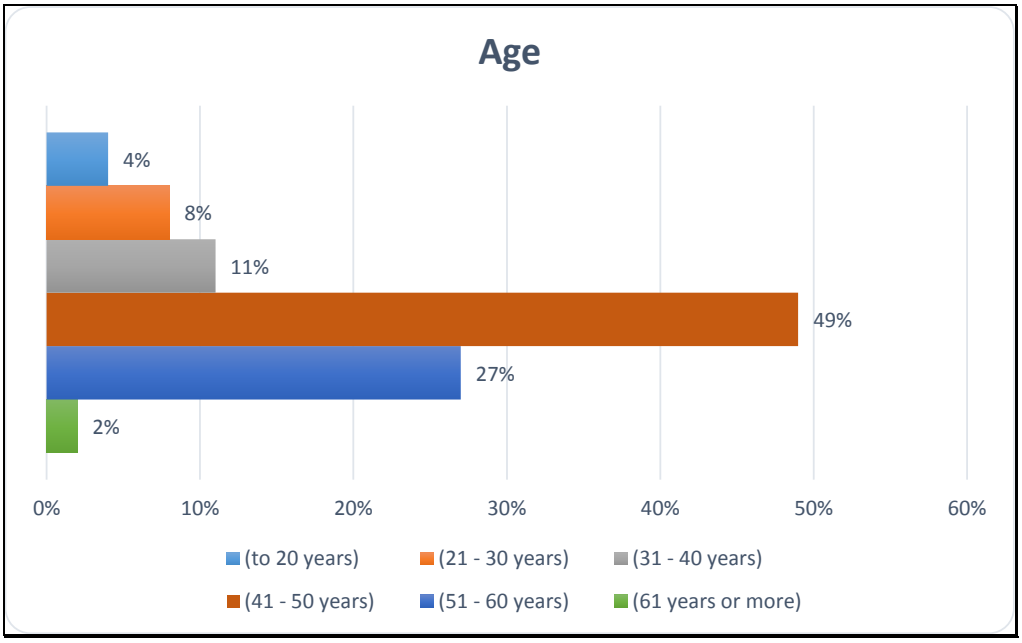


Figure 1 The age structure of participants in the sample.

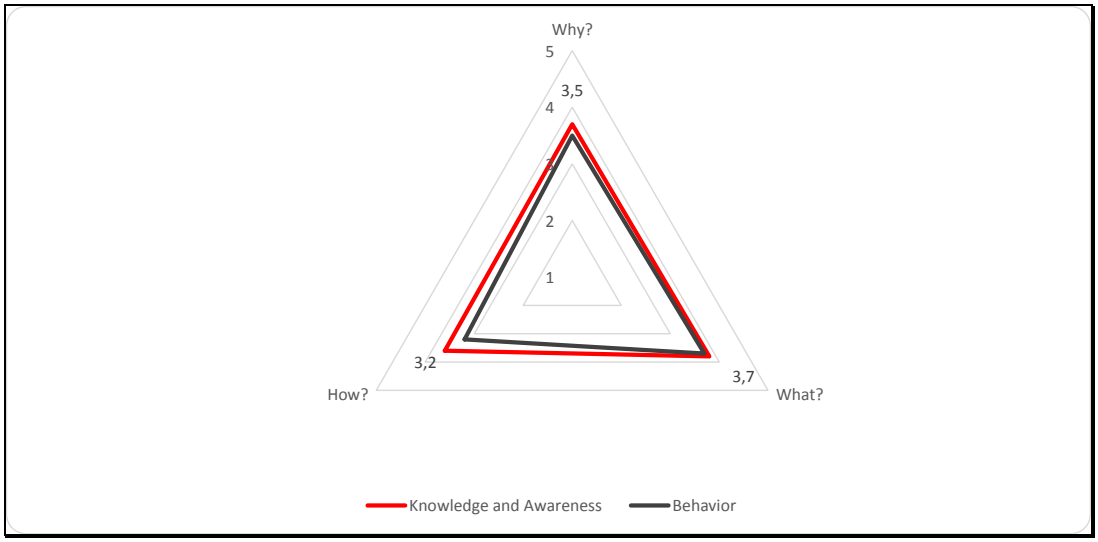


Figure 2 Gap between the knowledge and behaviour of online banking users.

Participants came from different regions. Among them, 65 were from Dolenjska and 51 from Bela Krajina.

Table 3 shows the results of individual concepts from the model and the Fig. 2 presents the gap between the user knowledge and behaviour. The validity of hypothesis was tested with the Z-statistics (comparing average values) and $\alpha = 0{,}01$.

Table 3 Average values of individual security culture components.

Information Security culture levels	User knowledge and perception of security	User behaviour, values and assumptions
Artefacts (How?)	3,6	3,5
Values (What?)	3,8	3,7
Assumptions (Why?)	3,7	3,2

The study results show that the information security culture of online banking users in Dolenjska in Bela Krajina does not reach the suitable level of 80% (less than 4 on the Likert scale). The knowledge part of the model shows better results but even there none of the components reaches the value 4. On average, the level of users' knowledge reaches 74%, while the level of behaviour is at 69%. We can conclude that users know more about security of online banking, but they behave worse. There is still quite a big gap in expected knowledge (6%) as well as in behaviour (11%). Apparently, the society has not done enough for sufficient safety within online banking users.

By comparing results in both observed regions, we cannot accept the hypothesis that the information security culture differs among users from Dolenjska and Bela Krajina. This hypothesis was placed due to our assumption that people from rural regions are less educated in IT and its security issues. It proved that this is not the factor, which would influence the information security culture.

6 Conclusion

This research was focused on information security culture of online banking users. As it is one of the most widely used IT services today and due to its attractiveness for the intruders, it is important for the culture of usage to be high. We analysed the level of security knowledge of users and their behaviour when using online banking. For this purpose, we proposed a measuring model. Results prove that the information security culture is still not sufficiently high. There is still some lack of knowledge about security, and besides that, even if they have knowledge, they do not behave safely.

In future work we are going to improve the measurement instrument in the way, that we will measure individual concepts of model with more questions.

7 Acknowledgements

Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

8 References

Da Veiga, A., & Eloff, J. H. P. (2010). A framework and assessment instrument for information security culture. *Computers & Security*, *29*(2), 196–207. doi:10.1016/j.cose.2009.09.002

Ella Kolkowska. (2011). Security subcultures in an organization - exploring value conflicts. In *19th European Conference on Information Systems, ECIS 2011, Helsinki, Finland, June 9-11, 2011*.

Mitnick, K. D., & Simon, W. L. (2002). *The Art of Deception: Controlling the Human Element of Security* (1st ed.). New York, NY, USA: John Wiley & Sons, Inc.

Niekerk, J. Van, & Solms, R. von. (2006). UNDERSTANDING INFORMATION SECURITY CULTURE. In *Proceedings of the ISSA 2006 from Insight to Foresight Conference*.

Okere, I., Niekerk, J. Van, & Carroll, M. (2012). Assessing information security culture: A critical analysis of current approaches. *Information Security for South Africa (ISSA)*.

Rančigaj, K., & Lobnikar, B. (2012). Vedenjski vidiki zagotavljanja informacijske varnosti : pomen upravljanja informacijske varnostne kulture, 1–12.

Ruighaver, a. B., Maynard, S. B., & Chang, S. (2007). Organisational security culture: Extending the end-user perspective. *Computers & Security*, *26*(1), 56–62. doi:10.1016/j.cose.2006.10.008

SOFTWARE DEFINED NETWORKS: AN OVERVIEW AND A SECURITY FEATURE PROPOSAL

Valter Popeškić
dr.sc. Božidar Kovačić
University of Rijeka
Department of Informatics, Rijeka, Croatia
{vpopeskic}@uniri.hr

Abstract: *This work can be described as a progression review document that follows last year's strictly theoretic view of cognitive network proposals and techniques in "Cognitive networks the networks of the future" article [1]. In the timeframe of one year after the presentation of cognitive network theory overview there are more than a few news in the field. Basically all of them are from the department of network device control plane centralization and evolution to real products this year.*
The buzzword SDN – Software defined networks took place from other very popular words like Virtualization and Cloud. Main goal of this work is to present a short overview of newly featured ideas, protocols and products from the field. The overview of existing technology is then enhanced with a proposal on possible future improvements that are in main focus of described research.
The work will comprehend all actualities in the networking field and it will correlate suggested automated network management methods with network security, network virtualization and few possible improvements of the whole plethora which may run network of tomorrow.

Key Words*: Software Defined Network, NFV - Network Feature Virtualization, BGP AS-PATH prepending, metric*

1 INTRODUCTION

A cognitive network is a network consisting of elements that reason and have the ability to learn. In this way they self-adjust according to different unpredictable network conditions in order to optimize data transmission performance. In a cognitive network, judgments are made to meet the requirements of the network as an entire system, rather than the individual network components. The main reason of the emergence of cognitive networks is to achieve the goal of building intelligent self-adjustable networks and in the same time improve the performance. Intelligent self-adjustable networks will be able to use measurement of network state and different probes results, convert all into statistic data to determine ideal network operating state for many tunable parameters.

2 NEW TYPE OF NETWORKS

2.1 Network planes and their function

To be able to further explain the network devices configuration centralization it is very important to list all different parts of common networking. The word that will emerge first is a plane. We can say that the plane is a networking context describing three components of networking function. Three different planes in networking are control plane, data plane and management plane. Each plane carries a different type of traffic. Data plane carries user traffic and its often called forwarding plane. Control plane is plane carrying signaling traffic and used by protocols that are running on network devices to synchronize their functions. At the end management plane is carrying administrative traffic and is basically a subset of the control plane. We are able to configure particular device using management plane. That device when configured is able to calculate and make routing decision using his routing protocol which in turn uses his control plane to communicate with the same routing protocol on other device in network. At the end device will forward received traffic based on the decision made in control plane that are more or less caused by our settings done in management plane.

Today networks are distributed. It means that devices mostly used in real world have all planes implemented on each network device. It means that management and every other function of a network are distributed. Every device is manageable locally. For example, routing process runs on every router in the network and based on information received from his neighbors that router chooses where to forward traffic. Note that information about best path is processed locally in operating system.

2.2 Software defined networks - SDN

"SDN is the physical separation of the network control plane from the forwarding plane, and where a control plane controls several devices." [2]

Open Networking Foundation

To be able to get you some insights into the future of network technology in this part we will shortly describe different terms of the software defined networking world. As stated early in this work network devices functionality is comprised of three planes of operation Management Plane, Control Plane and Forwarding or Data Plane. In the old fashion mode all three planes are part of every networking device.

The idea behind modern proposed software driven devices is that they do not have completely distributed management and control plane on every unit. The idea behind the SDN implementations is to have centralized SDN controller that will enable centralized control over configuration. All tough it sounds reasonable, it will be better to state that SDN is comprised of few SDN controllers because like any other networking system it will need to have some sort of redundancy[1] implemented in order to have stability, resiliency and even better scaling of the system.

[1] Redundancy - making duplicate critical components and functions inside a system in order to increase reliability of the system.

Each SDN controller in that case is able to connect to network devices and collect the network state details from which he will make a network model. From this model controllers will be able to show to the user the network state using specific user interface and it will be able to collect user commands that will make him generate and apply changes to network devices in network.

2.3 Most advantageous SDN feature - Automation

The best thing about SDN is not the ability to configure all devices from one centralized place. Though this is also a good thing. The best thing is that SDN controller can do it automatically. SDN controller has the ability to adapt making his own decisions and changing the configuration based on the end-to-end visibility of current network state. Controller is making decisions in a way that he computes the flow paths in his software. That is the main reason for having the name SDN – Software Defined Network relating to this new networking concept.

This kind of functionality will enable us to have more than just destination based hop-by-hop layer 3 forwarding that we do today. Maybe we will finally see easy to manage large-scale orchestration and provisioning toolset that will enable us to do security and QoS policies on different places in the network that may be dynamically tuned to current networking traffic state. Maybe even dynamically adjusted policy-based routing.

2.4 Misconception regarding SDN

SDN is regularly misinterpreted. There are more that few sources stating that SDN is all about centralizing the control plane. Nevertheless, SDN is a concept that is defining how to centralize the configuration for all devices in the data network system.

Decisions and forwarding is still done in distributed fashion and it is the way it should be in the future. The main goal of SDN concept is strictly defined. That is, enabling us to configure all devices from centralized place and have a mean to enable automatic tuning of network parameters based on current network state. In that way we must not connect to all 500 different devices in order to configure how they will forward the traffic afterwards or, for example, to tune some routing path to avoid congestion. The controller itself who is in charge to "control" and configure our 500 device needs to be distributed into several segments or places across datacenter in order to provide redundancy and make the system more scalable. You would then connect to one of those controllers and make some changes. That controller will then sync with other controllers and that push the configuration to devices where is needed. The most important thing to emphasize here is that there will be no centralized networking system. The system would need to stay distributed because that is one of the main characteristic of the networking after all.

2.5 Changing how network context are functioning

SDN in theory separates Data plane and Control plane by centralizing Control Plane. Like the definition on Software defined network from Open Networking Foundation it is described in short that we want to take control plane out of every device and centralize it on some sort of controller. The idea behind this is proven to be a good one. Centralization of control plane together with simple tough robust protocol (OpenFlow[2]) used to send the

[2] OpenFlow is a communications protocol that gives access to the forwarding plane of a network switch or router over the network [3].

configuration change information to each network device could derive with a stable and highly automated network system.

2.6 OpenFlow

SDN controller will use OpenFlow protocol to send changes to forwarding table of every network device. OpenFlow is a low-level tool bringing us a new way of controlling the configuration of forwarding table in the network switch from central location. It is basically a protocol with implemented API that is used by network SDN controller to configure network device [3]. So SDN controller has a user interface where users can configure some network component and then the controller is sending the configuration to the device using OpenFlow protocol.

2.7 Network Function Virtualization - NFV

There are several advantages coming with virtualized network management. One of them is surely cutting the cost of equipment. In future, there is a big chance that you will be able to buy very cheap not complex hardware for your network that supports OpenFlow. Of course there would always be a possibility to get very complex, feature rich devices running full IOS[3] or JUNOS[4] firmware but also controllable with OpenFlow. But that is only the management virtualization.

Network function virtualization also contains firewalls and load balancers, SIP gateways and Network appliances of different kinds. NFV basically means putting network services on a generic x86 hardware. The thing to point here is that NFV means also bare-metal installation of network services and operating systems not only virtualizing them on some sort of hypervisor[5]. The idea is to run those services with generic hardware skipping the one vendor appliances.

Virtualizing of network devices and their services is allowing us to simply move the network part together with applications that are running on our servers. This process is facilitating disaster recovery and cloud migration. Load balancers and firewalls will follow the application when you move it somewhere else in the datacenter, to other datacenter or to cloud.

2.8 Virtual firewalls today

Virtualization of firewalls is one of the most interesting parts of the whole NFV process. In the same time, it is the most controversial one because the security, stability and performance implications. This work will describe some of the most significant characteristics of today's virtualized firewalls with mostly performance and security in the focus. Later chapter is specially intended to give some ideas of security measurement and security improvement inside virtual network segment.

---

[3] IOS – Cisco IOS - Internetwork Operating System is software used on Cisco switches and routers.

[4] JUNOS – Juniper JUNOS - Juniper Junos is the FreeBSD-based operating system used in Juniper Networks routers and switches.

[5] Hypervisor - Virtual machine monitor (VMM) is a software, firmware or hardware that creates and runs virtual machines.

Most networking devices running layer 4[6] to layer 7 functions today are running x86 CPU. As from this, they all have the possibility to virtualize those appliances and offer them as VM[7] appliances that are running inside particular virtual machine.

Some of the firewalls that exist today are Vyatta from Broocade and vShield Edge from VMware. Other VM available products can be best described as load balancers like BIG-IP VTM or Zeus Traffic Manager and couple of start-ups that are doing similar products in creative way (Embrane, LineRate Systems).

First reason to virtualize firewalls is the issue with physical appliances regarding the inability to move physical device. This is particularly difficult for firewalls. Virtual firewall enables us that when we migrate the whole VM to other host, virtual firewall migrates with the workload and together with VMs. Of course, this would be only possible if we have separate virtual firewall for each tenant, each VM. The best part of last sentence is that it maybe sounds like something that is negative about virtual firewall, but is actually really positive and best case scenario. We are virtualizing firewalls because we want to move them together with the whole VM server but we also want to have a separate firewall in front of every server. This kind of implementation is per se futuristic but it would enable us to have granularity and simplicity in security management and firewall configuration. We could, in this way, have a template for newly created VM that will be able to create separate instance of virtual firewall put preconfigured rules on that firewall and attach it to the newly provisioned VM. Deployment and provisioning flexibility enables us to deploy per tenant/per application firewall that have first time configuration from template and are much easier to configure and furthermore to manage later in the process.

From here we see that some of the most obvious advantages are transport network independence, easy configuration management, workload mobility and simple and automatic deployment and provisioning.
From other side the main objections are focused on performance, attacks on hypervisors and multi-tenant attacks. Here we need to point out that the drawback is mostly about packet processing in CPU that is very expensive from every point of view, particularly for throughput. The main performance issues are related to mostly used Linux TCP/IP stack in virtual appliances and hypervisor virtual switch implementations that are adding their own I/O CPU processed overhead.

Those issues are more and more reduced using different techniques. There was different testing done on the TCP/IP stack performance on firewall VMs and the result from one of those was that replacing the Linux TCP/IP stack with proprietary one increases performance from 3Gbps to 18Gbps. The process includes connecting the VM directly to physical interface card. If you do not tie the VM directly to physical interface you lose the ability to use optimization techniques that TCP/IP stack has because you are going through hypervisor virtual switch. Other method to increase the throughput is by using TCP offload mechanisms. TCP offload is basically a technique where you send a large TCP segment to virtual switch and virtual switch just forwards that whole segment directly to physical network interface card. The interface card does the slicing of that data

[6] Layer 4 – OSI model Open Systems Interconnection model is a theory model that standardizes the internal functions of communication network system by partitioning it into abstraction layers.

[7] VM – Virtual Machine

into TCP segments that are sent across the network skipping the hypervisor process overhead. It shows performance increase from 3Gbps to 10Gbps all this if normal Linux TCP/IP stack was used. TCP offload for now is able to handle VLAN tagged traffic and in the next generation of network interface cards it is promised that there would be also support for VXLAN.

Bypassing the TCP stack by bypassing the kernel and in this manner allowing the process to directly access memory where the packet buffers are, is giving great throughput results up to 40Gbps through one Xeon server.

3 SECURITY FEATURE PROPOSAL

3.1 Introduction

Foremost reason for going down the road of networking virtualization in this work was a challenge to give a measure to one important mechanism that is mostly left out of the virtualized networking layer, security.

The main goal and focus of this work is the measurement of network path security. Be that in the virtualized and physical networking environment or in today's mixed environment. In conditions described here we need to take into account the virtualized part of the network communication path as the most difficult part to monitor and with that the most difficult part to protect and secure the communication that is going through.

Main goal of this chapter was to solve the main challenge in the process of data path security measurement – Data path selection. We needed to get some traffic engineering solution which will enable us to send the probes across different, mostly not best paths, directed towards specific destination. In order to get this to work we decided to go with BGP outbound path selection influencing mostly with AS-PATH[8] prepending.

Creating the experiment and studying methods of path selection on "normal" non-virtualized network segments will help us later bring that knowledge inside virtualized networking environment. Starting the experiment with BGP and AS-PATH prepending was a natural and simple way to get the results with path selection. Path selection will then be the starting point for making the system of probes which will test actual paths in virtual and non-virtual environment.

Taking this into account future development of the mentioned method, the main reason for this metric development will be its usage in the path selection mechanisms. That metric could then be incorporated in future versions of routing protocols and involved in metric fine tuning with current robust routing protocols. Giving the more concise use case it would help calculate real time security metric that will enable our routing protocols to select most secure path for every communication across the network that needs this kind

[8] AS-PATH Prepending procedure applies only to eBGP sessions—that is, when advertising prefixes to another AS and the local AS number is prepended in front of the AS_PATH attribute the number of times specified. BGP is always selecting the prefix with the shortest AS_PATH. The length of this attribute is probably the best approximation of the classic IGP metric when mapping this concept to BGP. This could be directly compared to the hop count concept used in RIP. Using this BGP attribute we can control inbound traffic path selection as one of most interesting traffic engineering technique inside BGP routing protocol.

of communication quality [4, 5, 6, 7]. In virtualized world it will enable us to select different virtual paths in VM intercommunication giving the possibility to select one path for secure communication of user applications maybe even across virtual firewall in between VMs and direct, completely "free" L2 path between those same VMs in non-private system communication would be needed.

3.2 Self-estimating the need to enforce the usage of network path security metric

Security metric implementation considers that it will be set-up to fall-over to standard routing whenever there will not be the need to have a secure communication channel. This can be the case of video and audio streaming in multimedia serving where the speed of the transfer and jitter control is of far greater importance than security of the transfer. Maybe there will be a suggestion to apply this security policing only to TCP and not UDP traffic. For the details about the concrete final implementation of the metric more detailed testing and simulations need to be done. Simulating different real life usage situations will enable us to learn more about what would be the best implementation scenario.

3.3 Idea

Idea about security metric starts to be built on the basis of standard routing protocols. In the beginning we decided to use BGP – Border Gateway Protocol implemented for IPv6 network. Multiprotocol BGP for IPv6 was the best candidate for the beginning of the experimentation giving the ability to transfer different information inside extension community attribute. Other suggestions and ideas included the usage of IPv6 RH0 headers. Additional IPv6 RH0 extension header have the ability to select waypoints towards destination for each packet sent. IPv6 extension header space is practically unlimited in size by using data payload space for expanding. From our calculation it can be used to insert up to 90 waypoint IPv6 addresses inside every packet. Using this two methods, layer 3 (IPv6) and layer 7 (MBGP) respectively enable us to have flexibility in measuring and applying path security metrics.

In routing process we have always multiple paths to get to each destination. There are surely more routes existing in the process of searching the way to get the packet to their destination. Those routes will be processed inside the router and inserted into RIB[9] (Router Information Base) but only the best one will be selected to be inserted into FIB[10] (Forwarding Information Base). Route inserted into FIB will be used by that device to forward traffic to next device toward destination. The question that we are asking is if that path that we are inserting into FIB is also the most secure one. If we want to determine that we need to use one more test before selecting which RIB route is going to be inserted into FIB. For testing purposes, in order to solve the issue of testing the paths that are not yet in the FIB we used IPv6 RH0 headers. IPv6 has the ability to get packets sent across specific intermediary next-hop addresses before going to the actual destination. In this way we can force test datagrams to cross different patch and calculate which of them was received on the other side with less or no errors in transmission. There are different ways to determine if a datagram has experienced attacks while crossing the path towards the destination. In that way we will be able to determine if the path in that specific moment has some attacker activity going on. Future experimentations gave the result that usage of RH0 headers in IPv6 is only theoretically an excellent choice. In the first phase of testing it was clear that today's networking equipment is more than half times configured

[9] RIB – Router Information Base

[10] FIB – Forwarding Information Base

with source based routing disabled by default. It will basically drop all of our packets with RH0 headers defined. Next step was to make the experiment using BGP and AS-PATH prepending traffic engineering technique.

4 EXPERIMENT AND ANALYSIS

Experiment was done using 90 Virtual Quagga routers running Zebra OS. Those routers run a specially written routing daemons that enables BGP, OSPF and other routing protocols on them. Protocols are implemented by the RFC standards. The fact that the solution is open source enabled us to get the experiment to the next level. Not only Unix based Quagga routers enabled us to run a simpler version of IPv6 Internet network but also to have the possibility to influence and change the mechanics inside routing protocols by different means and directly inside router OS [8]. That would not be impossible if using some vendor specific hardware that does not enable the owner to get the control over the internals of operating system.

The experiment was done in next few steps. At first we made the eBGP peering between every router and his neighboring devices making more than 300 connections peering between 90 devices. This way we had a situation of real world Internet customers and ISP interconnection scenario inside our virtualized environment.

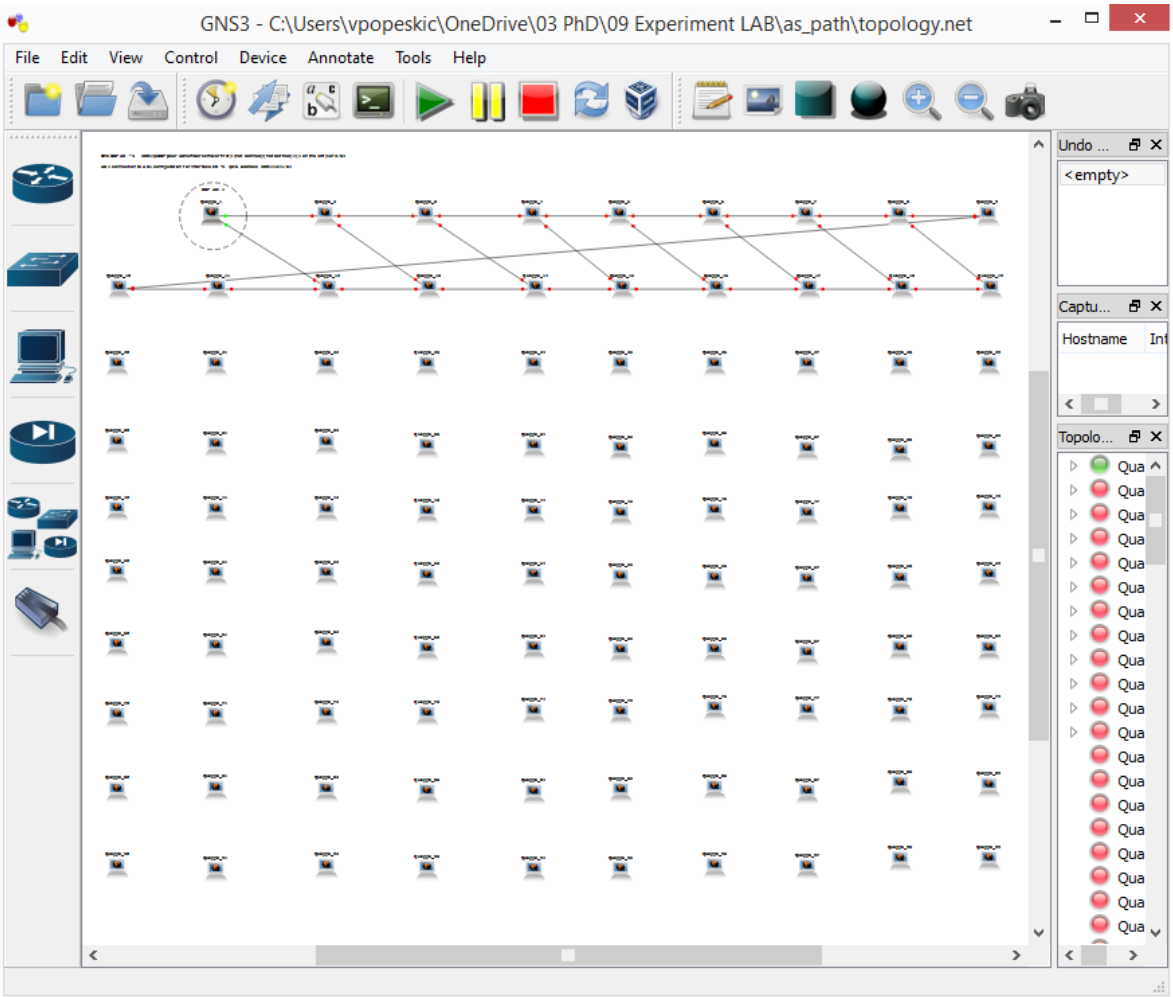


Figure 1, Virtual Quagga routers running Zebra OS simulating Internet BGP network for AS-PATH prepending experiment.

Be selecting all different AS numbers on the devices every device was basically a representation of the whole customer, whole ISP. It was simple enough for our purposes.

After making the connections and configuration of IPv6 addressing and BGP peering for all devices we continue to the main part of our experiment.

One router from environment edge was selected as the source of the testing and another from the other part of this environment was playing the role of destination.

We advertised the first /48 prefix out of our router to the "Internet" and after a short period of time all routers did learn the prefix and decided what will the best path to that prefix be across the network. Our source router did learn about other networks that other ASs have across the whole network. After this short period we could look at the BGP routing table inside our source router and read out all AS numbers that exist on the network. We also have a possibility to see which way our router decided to send the packets when they are forwarded to our selected destination device.

When we look at the route to our destination device on source router and read out the information about all AS numbers that are interconnections towards destination.

Next step was to subnet our /48 prefix to large number of /128 host networks. There is the possibility to get and be able to advertise about $1.2*10^{24}$ different host prefixes.
We selected the AS-PATH prepending method so that for every of those host prefixes we can advertise one or more AS-PATHs inside community header. If our destination prefix route in BGP showed 10 different interconnecting AS numbers we can advertise different /32 routes with every permutation of those AS numbers to get different paths back to our subnet. This method of traffic engineering is then used to test every of those paths in order to distinguish the path with best security metric between them. After the test, we can use that path for the whole subnet and advertise the /48 prefix with that combination of AS-PATH AS numbers prepended.

Experiment showed us that idea about sending data packets across different paths toward same destination was possible. It proved our idea about bypassing the best-path BGP route selection mechanism was also possible. It will enable us to test non-best paths regarding BGP for possibly better performance than best-path selected by BGP. BGP metric is complex but is unable to select the path based on real time security measurement or any other known performance metrics.

This could lead us to a new and reactive routing protocol feature that could test different routing paths before selecting the best one and thus have the ability to circumvent network congestions. Congestions of this kind are today a major issue in US. Having high quality multimedia streaming services in increasing number of countries across the world will also increase congestions on ISP interconnection peers. Having a reactive BGP metric controlled by SDN controller of some kind that could react upon congestion threshold would surely give to SDN one more reason to exist.

5 CONLUSION

This work is written to be an early insight of the new virtualization technologies trend called SDN and NFV. Furthermore, there was considerable effort invested in testing of the idea that it is possible to have better way to calculate best path for data traffic using into consideration real-time probing of secondary waypoint paths. SDN is a new way to implement our idea is more flexible and responsive way.

SDN as a technology is not new but it was long time trapped only in theory. Actually there was a considerable delay in network device virtualization considering the impact of virtualization technology on other IT departments and their existing devices like servers etc. Virtual networking devices were not in the process of becoming the reality for a long time. The emerging virtualization technology in data centres with millions of virtual server instances gave a big push to networking features virtualization development as well as development of centralized automated controllers to run those devices.

This article is not only about SDN controllers in future networking implementations. However it is true that tries to give an overview of this part of the networking technology. The main reason for this work is giving a ground for the future research in making network more secure. Analyses were made on network flows within networks of different kinds, virtual and physical. The idea that emerged was that there are some security measurements methods missing that would enable us to improve network path performance by a big percentage. After a period of research using different sources the idea becomes clearer every day. Perhaps this work will be a starting point for development of a part of SDN feature set that is yet to be made.

7 REFERENCES

[1] Popeškić, B. Kovačić, "Cognitive networks the network of the future." Society and Technology 2012 / Plenković, Mario (ur.). - Zagreb : HKD & NONACOM, 243-254 (ISBN: 978-953-6226-23-8), July 2012,

[2] https://www.opennetworking.org/sdn-resources/sdn-definition

[3] Nick McKeown et al. (April 2008). "OpenFlow: Enabling innovation in campus networks". ACM Communications Review. Retrieved 2009-11-02.

[4] Atef Abdelkefi, Yuming Jiang, Bjarne Emil Helvik, Gergely Biczók, Alexandru Calu, Assessing the service quality of an Internet path through end-to-end measurement, Computer Networks, Volume 70, 9 September 2014, Pages 30-44, ISSN 1389-1286

[5] N. Brownlee. Traffic Flow Measurement: Meter MIB.RFC2720. 1999

[6] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information theoretic approach to traffic matrix estimation. In SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, pages 301–312, Karlsruhe, Germany, August 2003.

[7] Myung-Sup Kim, et al, A flow-based method for abnormal network traffic detection. Network Operations and Management Symposium, 2004. IEEE/IFIP, Volume: 1, 19-23 April 2004 Pages:599 - 612 Vol.1

[8] Jakma, P.; Lamparter, D., "Introduction to the quagga routing suite," *Network, IEEE* , vol.28, no.2, pp.42,48, March-April 2014

[9] Will E. Leland, et al, On the self-similar nature of Ethernet traffic. IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 2, NO. 1, FEBRUARY 1994.

[10] K. Park and W. Willinger, Self-Similar Network Traffic: An Overview Self-Similar Network Traffic and Performance Evaluation. K. Park and W. Willinger (editors), John Wiley & Sons, New York, New York.

[11] Y. Vardi. Network tomography: estimating source-destination traffic intensities from link data. American Statistical Association, 91:365–377, 1996.

[12] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast accurate computation of large-scale ip traffic matrices from link loads. ACM SIGMETRICS, 31(1):206–217, 2003.

[13] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. Estimating point-to-point and point-to-multipoint traffic matrices: an information-theoretic approach. Sheng-Yuan Tseng, Yueh-Min Huang, Chang-Chun Lin, Genetic algorithm for delay- and degree-constrained multimedia broadcasting on overlay networks, Computer Communications, 2006, 29, 17, 3625

Game-based Learning and Social Media API in Higher Education

Petar Jurić
Department of Information Technology Services
Primorsko-goranska County
Adamićeva 10, 51000 Rijeka, Croatia
petar.juric@pgz.hr

Maja Matetić, Marija Brkić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{majam, mbrkic}@inf.uniri.hr

Abstract: *In the forthcoming years e-learning systems will be increasingly accessed by mobile computer devices, such as tablets and smartphones. Developing computer games for this platform will be a growing share of software development and of new learning methods. This paper gives analysis of motivational elements of computer games. An overview of the newest technologies for developing computer games in a mobile web environment and of the current research on using computer games for learning programming and for foreign language learning is given. A model on which our future work will be based is suggested. The model enhances the e-learning system of the University of Rijeka with computer game-based learning methods and the social network communication channel within the course Programming 2.*

Key Words*: motivation, computer game based learning, social media, m-learning, e-learning*

1 Introduction

Using motivational elements for knowledge acquisition and application is an important research area, which contributes to the learning outcomes. Motivation can be intrinsic (arises from within and it is reflected in the increased activity and the desire to participate evoking a feeling of satisfaction during the activity) and extrinsic (arises from other people's influence, e.g. parents', employers', teachers', etc.). Specific types of motivation such as integrative motivation might be present during foreign language learning in cases where mastery of a language is a precondition on community integration and communication. In instrumental motivation the learning desire arises from better job prospects, higher salary, etc., but does not necessarily evoke a feeling of satisfaction [13].

Current research aim at finding methods for increasing the level of motivation while learning supported by information technology. According to the Gartner's Hype Cycle for Emerging Technologies [6], the term gamification is identified at the peak of inflated expectations.

The concept gamification is originally found in video games. It refers to applying video game principles to everyday life or to applying elements found in games to

different products or services in order to encourage their use and add more pleasure when using them [5]. Applying the gamification concept primarily affects a person's emotions, or more precisely intrinsic motivation. Intrinsic motivation has stronger effects and lasts longer. It has a higher engagement level compared to extrinsic motivation [1].

The question that emerges is how computer games can be used in higher education with the goal of increasing students' motivation for new content acquisition.

2 Motivational elements in computer games

E-learning systems play an important role in present higher education systems, but lack of motivation for using those systems stands out as a main drawback. Newer generations of students (born in the last quarter of the twentieth century) exhibit different prevailing cognitive styles, which originate from computer games [14].

Motivational elements found in computer games are presented in Table 1. They are ordered by importance for the students. The ordering implies where attention should be primarily focused while processing educational content in computer games. The levels of action attached to these motivational elements may be individual or interpersonal.

Table 1: Motivational elements of computer games (adapted from [7])

Rank	Reasons	Description	Level
1st	Challenge	an appropriate level of difficulty and challenge, multiple goals for winning, constant feedback and sufficient randomness	individual
2nd	Curiosity	providing sensory stimulation	individual
3rd	Cooperation	assist others	interpersonal
4th	Competition	compare performance	interpersonal
5th	Control	the ability to select choices and observe the consequences	individual
6th	Recognition	a sense of satisfaction when accomplishments are recognized	interpersonal
7th	Fantasy	an appropriate level of immersion by assuming a particular role	individual

The research in this paper will be directed towards available technologies and toward the possibility of learning programming languages and foreign languages through computer games.

3 Open source technologies for programming computer games in a mobile web environment

According to the Recommendations for the Formulation of Educational E-Learning Material of the University of Rijeka [18] and the E-learning Development Strategy of the University of Rijeka [19], e-learning is based on the Internet (web) technologies. The goals listed in the Recommendations and the Strategy to which our future research will be directed are the following: 1. training by incorporating new technologies and applying new methods of learning and teaching which enable active knowledge

acquisition, and 2. opening possibilities for directing University activity towards new target groups of students by developing distance learning programs (courses).

The technology of Internet access and the way of consuming web content is changing due to the increase in the sales of smart mobile devices (tablets and smartphones) compared to desktop PCs and laptops. New devices have a touch screen as a primary input device (instead of a keyboard and a mouse) and different sensors which detect device orientation, direction and the angle at which the device is held, as well as device location and speed.

Smart mobile device applications can be developed in a native or a web-based environment. A short overview of native application development technologies compared to the web-based for the two dominant mobile platforms, Android and iOS, is given in Table 2, along with their pros and cons.

Table 2: Android and iOS native application development compared to the web-based development – pros and cons

Platform	Programming language	Development environment	Pros	Cons
Android	Java	Eclipse	Fast, a higher number of APIs for accessing hardware sensors	Platform-specific, changes are available only after reinstallation
iOS	Objective-C Swift	Xcode		

Android and iOS platforms are incompatible regarding applications developed for each one of them. Platform-specific application development means that a game created for Android needs to be developed in another development environment and another programming language for iOS [9].

From our point of view, the availability of e-learning content should not be platform-dependent, and it should be extendible and available on as many devices as possible. Since any changes in a web-based application are immediately available to all users, the speed benefit of running native applications, which is important for running games due to the high hardware requirements, falls into the background. Moreover, further fragmentation of mobile devices and technologies is restricted. By developing new APIs for HTML5 platform, the benefits of building native applications will be increasingly reduced [15].

In our future work, the development of educational games will, therefore, be based on: 1. open-source technologies, 2. platform and plug-in independent web environment, and 3. mobile device adaptation. The platform of choice will be the HTML5 ecosystem with responsive design implemented, which satisfies all of the three requirements. Currently, there is no framework of this type for educational game development. The HTML5 game development frameworks generally include collision detection and physical models of the real world, e.g. gravitation [8]. Game development is slow and, in most cases, frameworks do not have GUI, while only few have a level editor.

The familiarity with technology required of the teachers needs to be reduced to the lowest level for the wider acceptability of content processing through educational computer games. We suggest building predefined game types in which teachers could add textual and visual elements to be displayed on the learning objects in the game simply via administrator interface (Fig. 1). This module will form a central part related to learning new content, along with the integrated social media API, to which data mining systems will connect with the task of optimizing the learning process.

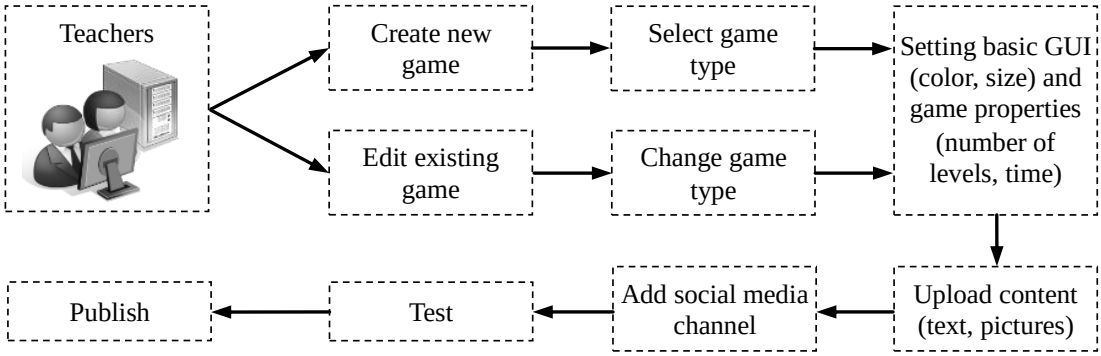


Figure 1: The process of creating educational computer games based on predefined types

4 Learning through computer games in higher education

Learning through computer games offers almost unlimited application possibilities in which real life situations can be mapped to or imaginatively displayed in a virtual world [3].

The work will focus on increasing motivation and on easier acquisition of more complex concepts within the course Programming 2 at the Department of Informatics of the University of Rijeka, i.e. concepts for which students obtain lower scores: 1. pointers, and 2. recursion. Similar research on improving learning of object-oriented programming techniques shows positive results [4].

Intersecting points between learning programming and learning other subjects, primarily English as a second language, will be identified. Common types of knowledge developed and acquired while learning programming and foreign languages are syntactic and conceptual knowledge [16]. Current research indicates that increasing motivation level has a greater impact on achievement than a predisposition for language acquisition [2].

A social media API will be implemented as a main textual communication channel between the students themselves and between the students and the teacher. Besides learning through computer games, using social media for thematic communication can also have a motivating effect [17].

5 The integrated model proposed and data mining in e-learning systems with computer games

Systems for learning through computer games can be attached to an e-learning system or they can be integral parts of it. The term m-learning or mobile learning is tied to smartphones. These systems can manage structured and unstructured data suitable for data mining [12]. Motivation and exam success can be predicted based on interaction with the game [10]. Structured data refer to location, speed, number of level attempts, time necessary to get to the next level, number of replays of the level after completing it, number of hints used, etc. Unstructured data originate in using integrated social media API either for communication or for seeking help from other students or the teacher.

The University of Rijeka uses MudRi learning management system (LMS), which is based on the open-source Moodle system.

The integrated model of a computer game based learning system (GBLS) with social network communication between users is given in Fig. 2. The data and text mining systems are attached to the model.

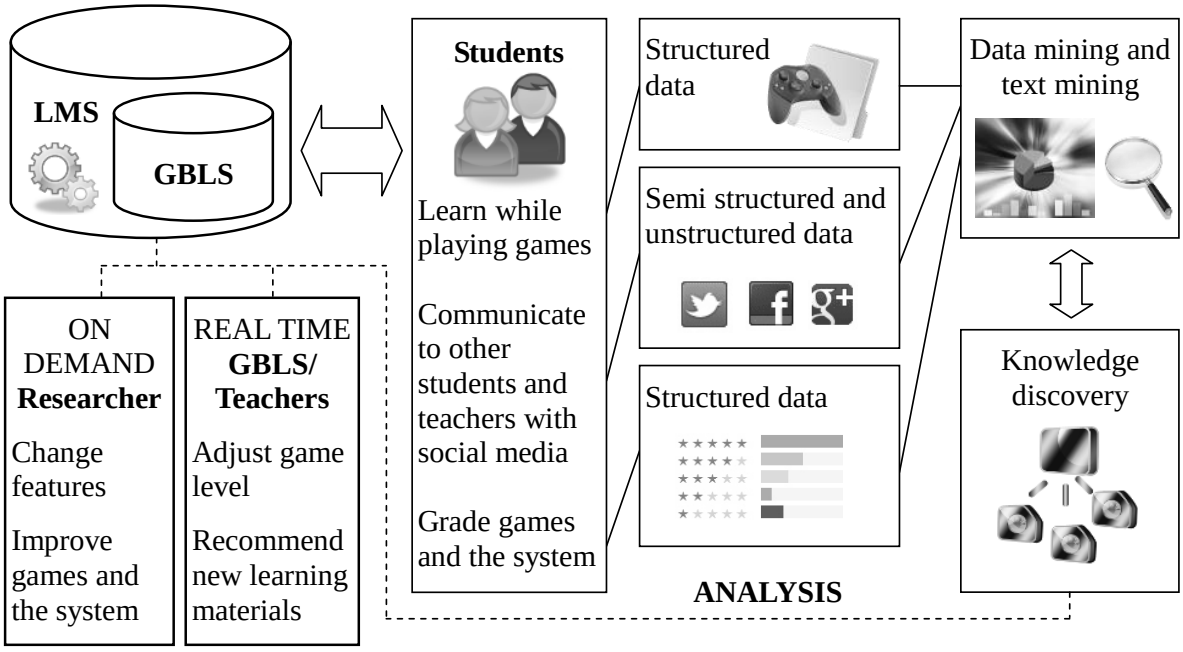


Figure 2: The integrated model of learning through computer games with social network communication channel and data mining within e-learning system

Data mining enables teachers to adapt game contents and game level to the level of students' prior knowledge and to recommend additional learning activities. It enables experts to analyze data in order to improve games by optimization, i.e. with the goal to achieve faster and more thorough approach to learning. Other data mining systems in the field that do not require expert knowledge might also be attached to the model [11].

For evaluating satisfaction with the learning approach a short questionnaire at certain points through the game or at the end of the game will be used, as well as game grading and social media API communication.

6 Conclusion

Using computer games in education triggers intrinsic motivational elements. Since the approach is fun, the time students spend learning, as well as the level of knowledge acquisition is increased. The reason why it is not widely represented lies in programming requirements and unfamiliarity with web technologies and technologies for game development. Besides complexity, another drawback is the time needed for the development. Since these elements have so far been developed in companies or by expert scientific teams, the proposed model which would enable developing games based on predefined types provides grounds for a simple and fast development of certain types of educational computer games.

By applying data mining techniques in these systems, the teachers could track the level of engagement and success prior to oral or written examination, as well as recommend materials or activities for remedial or additional classes.

7 Acknowledgement

This research has been supported under the Grant No. 13.13.1.3.03 of the University of Rijeka.

8 References

[1] Burke, B. Gamify: How Gamification Motivates People to Do Extraordinary Things. Biblimotion. Brookline, MA, USA, 2014.

[2] Daskalovska, N.; Koleva-Gudeva, L.; Ivanovska, B. Learner motivation and Interest. Procedia – Social and Behavioral Sciences, 46:1187-1191, 2012.

[3] De Freitas, S.; Rebolledo-Mendez, G.; Liarokapis, F.; Magoulas, G.; Poulovassilis, A. Developing an Evaluation Methodology for Immersive Learning Experiences in a Virtual World. In Proceedings of the Games and Virtual Worlds for Serious Applications VS-GAMES 2009, pages 43-50, Coventry, United Kingdom, 2009.

[4] Depradine, C. A. Using Gaming to Improve Advanced Programming Skills. Caribbean Teaching Scholar, 1(2):93-113, 2011.

[5] Deterding, S; Dixon, D.; Khaled, R; Nacke, L. From Game Design Elements to Gamefulness: Defining „Gamification“. In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pages 9-15, Tampere, Finland, 2011.

[6] Gartner. Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationship Between Humans and Machines, http://www.gartner.com/newsroom/id/2575515, downloaded Jun 21st 2014.

[7] Hainey, T.; Connolly, T.; Stansfield, M.; Boyle, E. The differences in motivations of online game players and offline game players: A combined analysis of three studies at higher education level. Computers & Education, 57(4):2197-2211, 2011.

[8] HTML5 Game Engines. Which HTML5 Game Engine is right for you? http://html5gameengine.com, downloaded Jun 27th 2014.

[9] IBM. Native, web or hybrid mobile-app development, http://public.dhe.ibm.com/software/in/events/softwareuniverse/resources/Native_web_or_hybrid_mobile-app_development.pdf, downloaded Jul 4th 2014.

[10] Illanas Vila, A.; Calvo-Ferrer, J. R.; Gallego Duran, F.; Llorens Largo, F. Predicting student performance in foreign languages with a serious game. In Proceedings of 7th International Technology, Education and Development Conference INTED2013, pages 52-59, Valencia, Espania, 2013.

[11] Jugo, I.; Kovačić, B.; Slavuj, V. A proposal for a web based educational data mining and visualization system. In Proceedings of the 5th International Conference on Information Technologies and Information Society ITIS 2013, pages 59-64, Dolenjske toplice, Slovenia, 2013.

[12] Jurić, P.; Matetić, M.; Brkić, M. Data Mining of Computer Game Assisted e/m-learning Systems in Higher Education. In Proceedings of the 37th International Convention MIPRO, pages 868-873, Opatija, Croatia, 2014.

[13] Ozgur, B; Griffiths, C. Second language motivation. Procedia – Social and Behavioral Sciences, 70:1109-1114, 2013.

[14] Prensky, M. The games generations: How learners have changed. Digital game-based learning, Paragon House, St. Paul, MN, USA, 2007.

[15] Puder, A.; Tillmann, N.; Moskal, M. Exposing Native Device APIs to Web Apps. First International Conference on Mobile Software Engineering and Systems MOBILESoft 2014, Hyderabad, India, 2014.

[16] Sanchez-Lozano, J. C. Exploratory Digital Games for Advanced Skills: Theory and Application. In Design and Implementation of Educational Games: Theoretical and Practical Perspectives, pages 92-107, IGI Global, 2010.

[17] Silius, K.; ... [et al]. Students' Motivations for Social Media Enhanced Studying and Learning. Knowledge Management & E-Learning: An International Journal, 2(1):51-67, 2010.

[18] The Committee for implementation of e-Learning at the University of Rijeka. Preporuke za izradu obrazovnih materijala za e-učenje [Recommendations for the Formulation of Educational E-Learning Material], http://www.uniri.hr/files/vijesti/Preporuke_e-ucenje_2009_UNIRI.pdf, downloaded Jun 22nd 2014.
[19] University of Rijeka. Strategija razvoja e-učenja na Sveučilištu u Rijeci 2011-2015 [E-learning Development Strategy 2011-2015], http://www.uniri.hr/files/staticki_dio/propisi_i_dokumenti/Strategija_e-ucenje_2011-2015.pdf, downloaded Jun 22nd 2014.

CONSEQUENCES OF IMPORTING NEGATIVE NEWS FROM ABROAD

Andrej Kovačič, Ph.d.
Faculty of Media
Leskoškova 9D , 1000 Ljubljana, Slovenia
{andrej.kovacic@ceos.si}

Nevenka Podgornik, Ph.d.
School of Advanced Social Studies in Nova Gorica
Gregorčičeva 19 , 5000 Nova Gorica, Slovenia
{nevenka.podgornik@fuds.si}

Abstract: *This article presents a detailed study of 2.606 randomly selected news headlines and short abstracts on the internet. Analyzing aggregated RSS news has enabled us to discuss the negativity of news reporting in a wider context. The main objective of this article is to answer an important research question: to what extent domestic and foreign ("imported") news from different media sources (television, newspaper and internet) evoke negative feelings. As Slovenia is a small country the lack of domestic negative shocking news is filled with foreign news, where the majority of them evoke negative feeling. Such disproportion consequently portrays a more negative image on the world outside Slovenia. This could lead to a misconception of foreigners and their culture and can seriously influence migration characteristics of Slovene people, which are one of the lowest in EU.*

Keywords: *RSS, news, foreign image, negativity*

Introduction and literature preview

A discussion about the consequences of negative news reporting has been abundant. Johnston and Davey [7] tested three groups with 14-min TV news bulletins that were edited to display either positive, neutral or negative material. Negative bulletin participants showed "increases in both anxious and sad mood", and a "significant increase in the tendency to catastrophize a personal worry". Similarly Johnston and Davey suggest that negative TV news programs can increase personal concerns that are not specifically relevant to the content of the program. What is of additional importance is that increases in negative mood as a result of viewing a negative news bulletin were also associated with increases in the catastrophizing of personal worries. Johnston and Davey also report that negative programs are likely to have a negative effect on mood by increasing individual's personal worries and anxieties. In fact, as Horn [6] explains, public discontent with the media in general results largely from exaggerated claims based on erroneous reasoning along these lines. Extreme consequence of negative news reporting can be found in Sudak H. and Sudak D. [17] as they provided examples how specific media reporting of suicide news has an impact on further suicides. Apollonio and Malone [1] analyzed negative political reporting and political advertising, another context of negative news reporting, and concluded that by increasing awareness to potential problems, media is actually encouraging people to reconsider their established opinions. Browne and Hamilton - Giachritsis [3] similarly analyzed violence in connection to the violence presented in media. They believe that long-term outcomes for children viewing media violence are controversial because of the methodological difficulties.

The relationship between mood (influenced by media) and long-term health risk is pointed out in a thirty-five-year longitudinal study by Peterson, Seligman and Vaillant [14]. This study indicates that pessimistic style predicted poor health at ages 45 through 60, even when physical and mental health at age 25 were controlled. Thus pessimism in early adulthood appears to be a risk factor for poor health in middle and late adulthood.

Slovenian media research is limited. A decade old research was done by Petrovec [15], who claims media reporting is focusing on the most spectacular stories mainly presenting violence in volume that is not representing the reality. In addition Petrovec claims negative news reporting is contributing to the idea of bringing fear in the Slovenian society. In his research almost 80 % of Slovenian people are convinced that Slovenian media is too full of programs containing violence and thus urge for a sort of a guardian of public interest in media reporting. In POP TV the share of news containing violence was 40.3 % and in TV SLO 1 program 18.2 % [17]. Similarly in a tabloid Slovenske novice the share of news containing violence was 26.9 %, whereas the share of words in headlines implying violence was incredibly over 80 %.

Heath [5] claims that achieving nor attention nor engagement means you are in a serious danger of having an unsuccessful message. "Bad" news attracts more attention than "good" news. The incentive to buy (purchase newspapers or spare television time) is so closely related to visual attention. The latter is defined as "the allocation or concentration of cognitive energy to process one part of the visual field at the expense of other parts" ([8], [4]). Young [18] refers to this objective as "stopping power". For advertising he states that in order to be effective "print advertisement must get noticed and attract a reader".

Research hypotheses, research design, method and sampling

We set the research question as to whether or not there is a difference in Slovenian and foreign news reporting tone. We hypothesize that a small country that does not have enough shocking domestic news tends to use foreign negative news to fill in the gap. This can lead to many misperceptions of the outside world and can influence perception and migration preferences.

We used a random sample for each media with each record (news) within the group having exactly the same probability of being selected without repetition. The number of evaluated articles (sample number) can be seen in table 1.

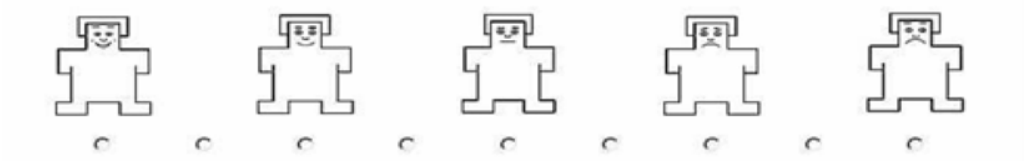
Table 1: Sample numbers for different media

Language group	Sample number	Domestic news topic	Foreign news topic
TV	651	305	346
Newspaper	977	569	408
Free newspaper	648	366	282
Internet	330	200	120
Total	2606	1166	1440

Source: own research 2013, 2014

This selection procedure assured the normal distribution as well as sufficient representation of each of the selected media. In the analysis the independent variable was the evaluation of the story based on a one item 9- point Likert scale a pleasure dimension of SAM (appendix 1) - "How did this news make you feel?" Self-Assessment Manikin (SAM) is a visual non-verbal scale with a graphic character arrayed along a continuous 9-point scale with Cronbach alpha= 0.82 and 0.98 ([2], [12], [13], [16], [17]).

Figure 1: How did reading the news make you feel?



1 2 3 4 5 6 7 8 9

Source: adapted from Moris [2], [12]

Results of the evaluation were calculated to range from -4 (extremely negative evoked feelings) through 0 (neutral) to 4 (extremely positive evoked feelings). Although Likert scale is in its essence ordinal in this example (SAM) can approximate an interval-level used for measurement ranging from -4 to +4. A total of 2606 news was evaluated, each news being evaluated twice to test the reliability of the evaluators.

Inter-rater reliability calculations

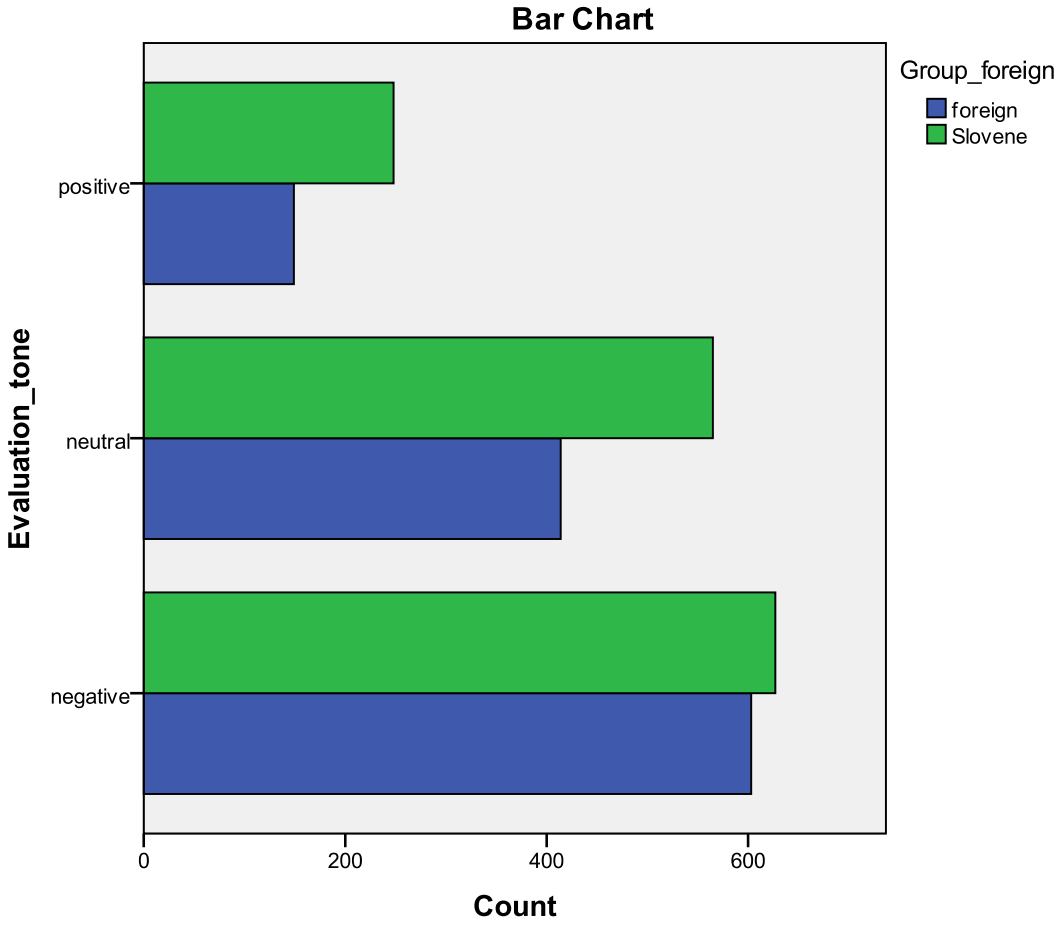
To improve the reliability of this research each news headline and short summary (up to 250 characters) was evaluated twice by two evaluators. The evaluators were independent and were rewarded in money for the task. Their judgments were not influenced by any of the authors or support team neither was any of the authors of this article a part of the evaluation team. Evaluators were native speakers living in Slovenia. Following the evaluation a Krippendorff's alpha [10] on interval variables was used to evaluate the inter-rater reliability. Krippendorff's alpha is exclusively rooted in the data generated by all observers and defines the two reliability scale points as 1.00 for perfect reliability and 0.00 for the absence of reliability which are statistically unrelated to the units they describe. The sampling distribution of the means is assumed to be normally distributed as well as the sampling distribution of the scores. Reliability calculations for media tone using Krippendorff's alpha showed alpha=0.926. Calculated reliability for evaluation on SAM showed alpha = 0.945. We also tested the reliability for the separation of foreign and domestic news and it showed alpha = 0.959. All calculated alphas are to be considered a reliable variable for analysis. Krippendorff [11] suggests relying on evaluators with variables $\alpha \geq 0.80$ although $\alpha \geq .667$ can suffice for drawing tentative conclusions. For example alpha 0.90 % means that 90 % of the units tested by evaluators are perfectly reliable while only 10% are the results of chance. In addition 55 news articles from different sources were compared using RSS feed and actual broadcasted/printed material and were thus tested to assure RSS feed is indeed the actual news reporting. All 55 articles were identical and thus we can conclude that RSS feed is indeed a precise reflection of actual media reporting.

Results

A One-Way Analysis of Variance (oneway ANOVA), with Post Hoc Comparisons, was used to analyze whether means for Slovenian and foreign news differ. The analysis of the independent variable was the evaluation of the story based on a one item 9- point Likert scale. The results of the evaluation were calculated to range from -4 (extremely negative evoked feelings) through 0 (neutral) to 4 (extremely positive evoked feelings).

The difference between foreign and Slovene news tone can be clearly seen on Figure 2. There are substantially less articles that are positive portraying foreign news.

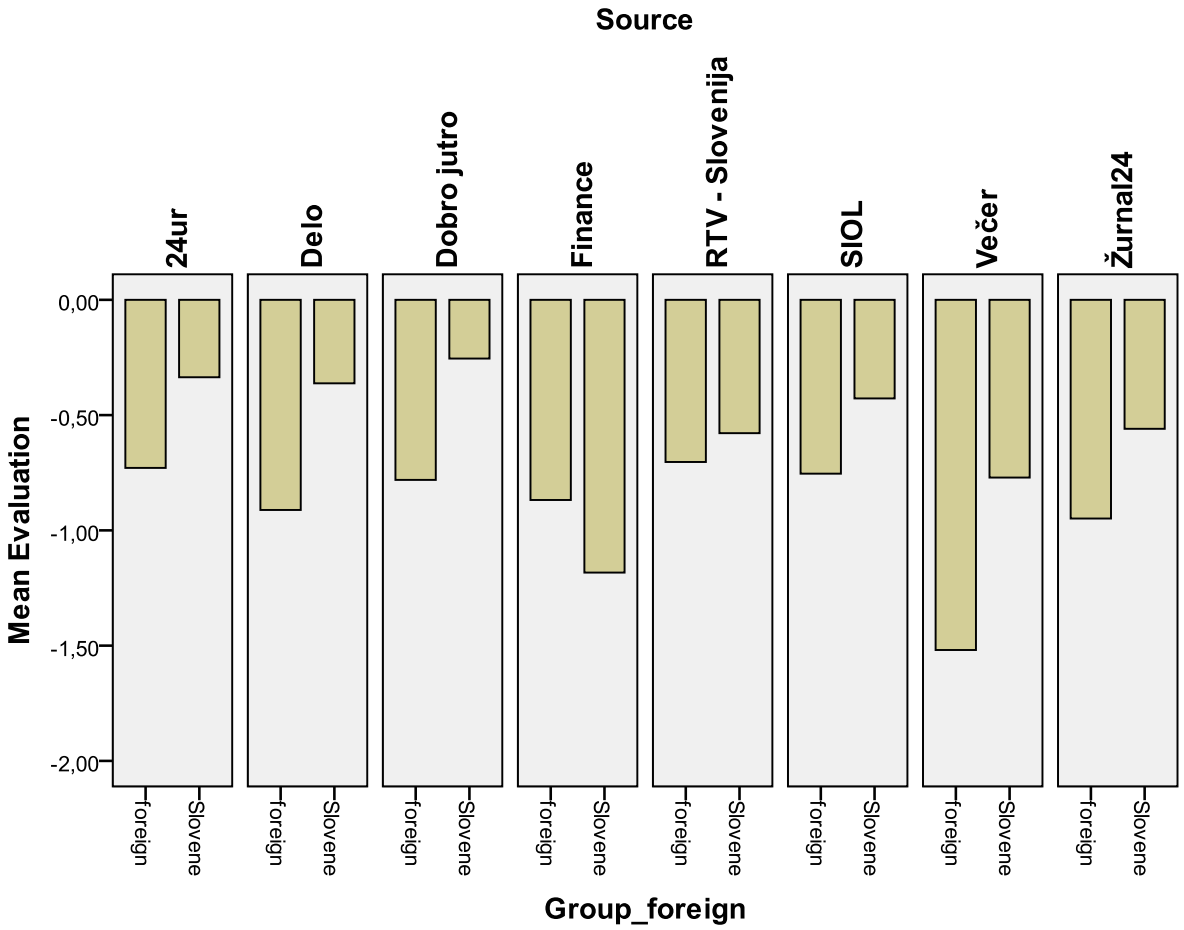
Figure 3: Difference between foreign and Slovene news reporting in a count of articles



Source: Own research 2013

The results from the analysis indicate that there is a significant difference between the two groups $F(3,2606) = 9.224$, $p < .001$. The mean values (-0,63) for the evaluations of the Slovenian news differ from evaluations of foreign news (-1,03).

Figure 2: Difference between the average foreign and Slovene news reporting evaluation in different media



Source: own research, 2013

Figure 3 shows us that 7 out of 8 media portray more negative foreign news than Slovene news. The only exception against the suggested hypothesis is Finance (a financial newspaper). Altogether media portrays on average from 0.1 to 075 points on a 9-point likert scale lower evaluation for foreign news (lower is more negative). Thus we can confirm our hypothesis that media indeed selects more negative foreign news and less positive foreign news.

Conclusions and implications

So much negative news especially from abroad is a broad and systemic problem of media reporting. This practice is stimulated by limited media attention and profitable media outcomes. Even more, by being able to change values and criteria of the people towards extremes of what is acceptable, the tendency is strongly leaning towards even more negative news reporting.

Media is strongly biased towards negative news reporting and as such (ab)uses foreign news to fill in the gap. What happens to the image of foreigners if you portray foreign news that are very negative? This research has proven that Slovenia as a small country imports a large number of foreign news that evoke negative feelings. This results in portraying a more negative image on the world outside Slovenia and might lead to a misconception of foreigners and their culture. To what extent negative news can influence migration characteristics of Slovene people is a complicated question and needs to be addressed in the future research.

Finding a solution to this serious issue of negative news reporting is difficult. On the one hand regulation can improve the ratio of the negativity in news reporting as recognized also by Horn [7]. On the other hand media is lobbying strongly for deregulation, promising existing politicians support on elections. Nevertheless we have to consider the results of this study to provide a strong incentive for future research and legislative action.

References

[1] Apollonio, D. E., & Malone, R. E. (39753): Turning negative into positive: public health mass media campaigns and negative advertising. Oxford: Health Education Research - Oxford University Press. Vol.: 24, N.: 3, pp.: 483-495.

[2] Backs, Richard W., Silva, Sergio P., & Han, Kyunghee (2005): A Comparison Of Younger And Older Adults' Self-Assessment Manikin Ratings Of Affective Pictures. Experimental Aging Research. Vol.: 31, pp.: 421-440.

[3] Browne, Kevin D; Hamilton-Giachritsis, Catherine. (2005): The influence of violent media on children and adolescents: a public-health approach. London: The Lancet (Elsevier Limited). Vol.: 365, N.: 9460, pp.: 702-710.

[4] Cummings, Maria N. (Apr, 2007): Consumer Engagement Perspectives: A Tool for Ensuring Advertising's Impact?.Rochester: School of Print Media Rochester Institute of Technology. Rochester, New York.

[5] Heath, Robert (2007): Emotional Persuasion in advertising: A Hierarchy-of-Processing Model.Bath: University of Bath.

[6] Horn, Karen (2007): A Market Like Any Other: Against the Double Standard in Judging the Media. Oakland: The Independent Review. Vol.: 12, N.: 1, pp.: 27-46.

[7] Johnston, Wendy M; Davey, Graham C. L. (1997): The psychological impact of negative TV news bulletins: The catastrophizing of personal worries. London: British Journal of Psychology. Vol.: 88, pp.: 85-91.

[8] Ketelaar, Paul E., Gisbergen, Mamix S. Van, Bosman, Jan A.M., & Beentjes, Hans (2008): Attention for Open and Closed Advertisements. Journal of Current Issues in Research in Advertising. Vol.: 25, No.: 2, pp.: 30-15.

[9] KOVAČIČ, ANDREJ (2011): How do Media Report News in Slovenia? Is there a negative Bias in Communication?. In Viera Žuborova, Diana Kamelia & Uroš Pinterič: Social responsibility in 21st century. Ljubljana: Vega.

[10] KRIPPENDORFF, KLAUS (2006): Testing the Reliability of Content Analysis Data: What is Involved and Why. Pennsylvania, ZDA: University of Pennsylvania. Available at: http://repository.upenn.edu/asc_papers/43 (12.5.2014)

[11] KRIPPENDORFF, KLAUS (2011): Computing Krippendorff's Alpha–Reliability. Pennsylvania, ZDA: University of Pennsylvania. Available at: http://repository.upenn.edu/asc_papers/43 (9.2.2014)

[12] Moris, Jon D. (1995): SAM: Self Assesement Manikin - an Effecient Cross-cultural Measurement of Emotional Response. Journal of Advertising. Vol.: 38, No.: 1, pp.: 123-136.

[13] Morris, Jon D., Woo Chongmoo, Geason, James A., & Kim Jooyoung (2002): The Power of Affect: Predicting Intention. Journal of Advertising Research., pp.: 7-17.

[14] Peterson, Christopher; Seligman, Martin E.; Vaillant, George E. (1988): Pessimistic explanatory style is a risk factor for physical illness: A thirty-five-year longitudinal study. Journal of Personality and Social Psychology. Vol.: 55, N.: 1, pp.: 23-27.

[15] Petrovec, Dragan (2003): Mediji in nasilje : obseg in vpliv nasilja v medijih v Sloveniji. Ljubljana: Mirovni inštitut.

[16] Poels, Karolien, & Dewitte, Siegfried (March, 2006): How to Capture the Heart? Reviewing 20 Years of Emotion Measurement in Advertising. Journal of Advertising Research., pp.: 18-37.

[17] Sudak, Howard S; Sudak, Donna M (2005): The Media and Suicide. Washington: Academic Psychiatry. Vol.: 29, N.: 5, pp.: 495-499.

[18] Young, Charles E. (2010): Print Ad Research. Ameritest/CY Research, Inc.Available at: http://www.ameritest.net (12.7.2011)

Inferring Structure of Complex Dynamical Systems with Equation Discovery

Zoran Levnajić[1], Ljupco Todorovski[2], Bernard Ženko[3]
(1) Faculty of Information Studies in Novo mesto
(2) University of Ljubljana, Faculty of Administration
(3) Jozef Stefan Institute, Department of Knowledge Technologies
zoran.levnajic@fis.unm.si, ljupco.todorovski@fu.uni-lj.si, bernard.zenko@ijs.si

Abstract: *Inference of the inner structure of complex dynamical systems from incomplete observational data is an omnipresent and realistic problem in various scientific fields. Interest in the structure of a complex system comes from the need to better understand its functioning and design principles. Results hereby presented fit in the framework of dynamical networks of nonlinear influences between nodes that correspond to the variables of an observed dynamical system. The inference task we address relates to determining the network structure that best reconstructs a dynamical behavior given through a time series of empirical measurements of the system variables. We address the inference task using machine learning methods for equation discovery. A particular method used here is ProBMoT; it learns process-based models from data and user-provided knowledge about modeling in the domain at hand. We encode knowledge about the nonlinear interactions between the network nodes. For generating training data, we select four simple dynamical networks, simulate their behavior and add various intensities of artificial Gaussian noise. We then use ProBMoT to infer the network structure from the training data and encoded knowledge. We evaluate the inferred structures by comparing them to the original network structures used to generate the training data. The analysis of the comparison results shows a high degree of match between the inferred and original network structures as well as modest method robustness to increasing noise intensity. Our approach has the potential of developing into a method able to infer structure even without knowing the details of the inter-node dynamical interactions.*

Key Words*:* *network inference, dynamical systems, machine learning, equation discovery…*

Modeling wireless networks using graph theory

Jaka Kranjc, Janez Povh, Borut Lužar
Faculty of Information Studies in Novo mesto
{jaka.kranjc, janez.povh, borut.luzar}@fis.unm.si

Abstract: *A wireless network consists of access points and clients, distributed in a plane. In order to exchange information, a data carrier, i.e. a frequency, is assigned to each access point. The task of assignment of frequencies is to minimize the area of interference. A typical model for such problems is a graph in which access points are represented by vertices and every two interfering access points are connected by an edge.*

In this talk we will present a solution of the above problem using a max-k-cut approach. Additionally, we will present a model of a problem of distributing access points in a building and a possible solving method.

Key Words*:* *wireless networks, frequency assignment…*

Extremal graphs with respect to vertex betweenness for certain graph families

J. Govorčin [1], R. Škrekovski [1,2,3]

October 31, 2014

[1] Faculty of Information Studies,
Ulica talcev 3, 8000 Novo mesto, Slovenia
jelena.govorcin@fis.unm.si

[2] Institute of Mathematics, Physics and Mechanics,
University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia
skrekovski@gmail.com

[3] Faculty of Mathematics, Natural Sciences
and Information Technologies, University of Primorska,
Glagoljaka 8, 6000 Koper, Slovenia

Abstract

In this paper, we consider the graph-theoretical concept of the betweenness centrality measure. Our attention is focused on the determination of the extremal betweenness values within the various families of graphs. We prove that the maximum betweenness value is reached for the maximum degree vertex of the fan graph $F_{1,n-1}$ (resp. the wheel graph W_n) in the class of 2-connected (resp. 3-connected) graphs. In addition, we study the family of graphs with prescribed maximum degree as well as the class of graphs with diameter at least 3. Moreover, the extremal graphs between all graphs of minimum degree at least 2 or 3 are found. At the end, we describe some illustrative computations with real-world networks.

Keywords: betweenness centrality, extremal value, degree, connectivity, diameter

1 Introduction

Identifying vertices having a key role within large-scale graphs is attracting an increasing interest with numerous applications in real-world networks. Measures of importance of objects within complex networks are formally expressed by so called *centrality indices*. According to the nature of relations between objects of the network and the criterion of importance, one can consider various centrality indices: degree, closeness, betweenness or eigenvector centrality.

As a centrality index, betweenness quantifies the number of times a vertex acts as a bridge along the shortest path between two other vertices. Due to the assumption that a flow (for example, transport or information flow) between vertices is propagated mainly along shortest paths of the network, the vertices that lie on many shortest paths can profit from the flow influence more than the vertices which are avoided by shortest paths.

Besides practical applications, an attention is recently paid also to the graph-theoretical properties of betweenness centrality [3, 4, 13], as well as to its connection to the mean distance in graphs [2, 7, 8]. On the other hand, much less is known about values that the betweenness of vertices within a graph can reach, even in the case when all values are the same rational number (see [9]). In [5], it was shown that the maximal value of the betweenness centrality measure within a graph is $\binom{n-1}{2}$ and it is attained if and only if the considered graph is isomorphic to the star, i.e. $K_{1,n-1}$.

In this paper, we consider the graph-theoretical concept of the betweenness centrality measure. More precisely, we are interested in determining the extremal betweenness values within the various families of graphs, such as the families of graphs with prescribed maximum or minimum degree, 2-connected and 3-connected graphs, as well as graphs with diameter at least three. In addition, we describe some illustrative computations with real-world networks and compare the real betweenness values with the obtained maximum values.

2 Definitions and notations

Before presenting the main results of our study, we give some relevant definitions and notations which underlie our work. Further information regarding the definitions, notations and families of graphs can be found in [1].

Throughout this paper, all graphs are assumed to be undirected, finite and connected, without loops or multiple edges. Given a graph G, we denote $n := |V(G)|$ the number of vertices and $m := |E(G)|$ the number of edges of G. A *path* between two vertices $u, v \in V(G)$, denoted (u, v)-path, is a sequence of vertices starting at u and ending at v, i.e. $us_1s_2 \cdots s_kv$, such that from each of its vertices there is an edge to the next vertex in the sequence. The *distance* between two vertices u and v in a graph G is the length of a shortest (u, v)-path and is denoted by $d(u, v)$. The *diameter* of a graph G, $\mathrm{diam}(G)$, is the maximum distance between any pair of vertices $u, v \in V(G)$.

The join $G = G_1 + G_2$ of graphs G_1 and G_2 with disjoint vertex sets $V(G_1)$ and $V(G_2)$ and edge sets $E(G_1)$ and $E(G_2)$ is the graph union $G_1 \cup G_2$ together with all the edges joining $V(G_1)$ and $V(G_2)$. The disjoint union of k copies of a graph H is denoted by kH. The *fan* $F_{r,s}$ is defined as the graph join $\overline{K}_r + P_s$, where $\overline{K}_r$ is the empty graph on r vertices and P_s is the path on s vertices. The *wheel* W_n of order n is the graph join $K_1 + C_{n-1}$, where K_1 is the singleton graph and C_{n-1} is the cycle on $n-1$ vertices. The *windmill* $\mathrm{Wd}\,(r, s)$ is the graph obtained by taking s copies of the complete graph K_r with a vertex in common. Thus, it has $sr - s + 1$ vertices.

The *betweenness centrality* $b_G(x)$ of a vertex $x \in V(G)$ is the relative number of shortest

paths between all pairs of vertices passing through x:

$$b_G(x) = \sum_{\substack{u,v \in V(G) \\ u \neq v \neq x}} \frac{\sigma^G_{u,v}(x)}{\sigma^G_{u,v}}, \tag{1}$$

where $\sigma^G_{u,v}$ denotes the total number of shortest (u,v)-paths in G and $\sigma^G_{u,v}(x)$ represents the number of shortest (u,v)-paths passing through x. The index G is omitted if G is known from the context. By $\bar{b}(G)$, we denote the average betweenness of a graph G, and by $b_{\min}(G)$ (resp. $b_{\max}(G)$) the smallest (resp. the largest) vertex betweenness within the graph G.

For a graph family $\mathcal{H}$ and an integer n, let

$$B_{\max}(\mathcal{H}, n) := \max\{b_{\max}(G) : G \in \mathcal{H}, |V(G)| = n\}.$$

Then, within the family of all graphs, denoted by $\mathcal{G}$, we have $B_{\max}(\mathcal{G}, n) = \binom{n-1}{2}$. In this paper, we consider the previously mentioned problem of maximization within several proper subfamilies of the family $\mathcal{G}$ – the family $\mathcal{G}^\Delta$ of all graphs with the maximum degree at most Δ, the family $\mathcal{G}_\delta$ of all graphs of minimum degree at least δ, the families $\mathcal{C}_2$ and $\mathcal{C}_3$ of all 2- and 3-connected graphs and the family $\mathcal{D}_D$ of graphs with diameter D. We present exact values and estimates for extremal values of maximum betweenness within graphs from previously defined families, together with extremal graphs realizing these values.

3 Extremal values and extremal graphs

In the proofs of our results, we often use the following way of calculating the betweenness of a vertex x within a graph G: let $\mathcal{V} = \binom{V(G)\setminus\{x\}}{2}$, $\mathcal{P}$ being the set of all pairs $\{u,v\}$ from $\mathcal{V}$ that form an edge in G, $\mathcal{Q}$ being the set of all pairs $\{u,v\}$ from $\mathcal{V} \setminus \mathcal{P}$ such that there is a 2-path uyv with $y \neq x$, and finally, $\mathcal{R} = \mathcal{V} \setminus (\mathcal{P} \cup \mathcal{Q})$. Then

$$b_G(x) = \sum_{\{u,v\} \in \mathcal{P}} \frac{\sigma_{u,v}(x)}{\sigma_{u,v}} + \sum_{\{u,v\} \in \mathcal{Q}} \frac{\sigma_{u,v}(x)}{\sigma_{u,v}} + \sum_{\{u,v\} \in \mathcal{R}} \frac{\sigma_{u,v}(x)}{\sigma_{u,v}}. \tag{2}$$

Note that in (2) the first sum always contributes 0. Each term in the second sum contributes at most $1/2$ and each term of the third sum contributes at most 1.

Now, let G be a graph on n vertices, w be a vertex of G of maximum betweenness and let $H = G - w$. Denote by $e_1(H)$ the number of adjacent pairs of vertices in H and by $e_2(H)$ the number of pairs of vertices at distance 2 within the graph H. Since the contribution of the adjacent pairs of vertices is 0 and pairs of vertices at distance 2 can contribute at most $1/2$, using (2) we obtain

$$b(w) \leq 0 \cdot e_1(H) + \frac{1}{2} \cdot e_2(H) + 1 \cdot \left(\binom{n-1}{2} - e_1(H) - e_2(H) \right).$$

The above expression can be simplified to

$$b(w) \leq \frac{(n-1)(n-2)}{2} - e_1(H) - \frac{1}{2} \cdot e_2(H). \tag{3}$$

In what follows, we present the most important results of our work. First, we consider graphs with prescribed maximum degree, thus all graphs from the family $\mathcal{G}^{\Delta}$.

Proposition 1. *(·, Hurajová, Madaras, Škrekovski)*

If G is a graph of maximum degree Δ, then

$$b_{\max}(G) \leq \frac{\Delta - 1}{2\Delta} (n-1)^2.$$

In the following two propositions, we consider the families of 2- and 3-connected graphs, i.e. families $\mathcal{C}_2$ and $\mathcal{C}_3$.

Proposition 2. *(·, Hurajová, Madaras, Škrekovski)*

If G is a 2-connected graph on n vertices, then

$$b_{\max}(G) \leq \frac{(n-3)^2}{2}.$$

Moreover, the bound is obtained only at the central vertex of the graph $F_{1,n-1}$.

Proposition 3. *(·, Hurajová, Madaras, Škrekovski)*

If G is a 3-connected graph on n vertices, then

$$b_{\max}(G) \leq \frac{(n-1)(n-5)}{2}.$$

Moreover the bound is obtained only at the central vertex of the wheel graph W_n.

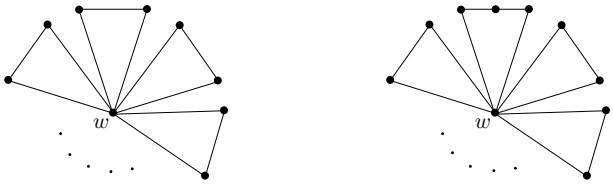
Now, we are interested in the family of graphs with prescribed minimum degree δ, which we denote by $\mathcal{G}_{\delta}$.

Proposition 4. *(·, Hurajová, Madaras, Škrekovski)*

If G is a graph on n vertices with minimum degree at least 2, then

$$b_{\max}(G) \leq \frac{(n-1)(n-3) - 1 + (-1)^{n+1}}{2}.$$

The maximum for n odd is attained for the central vertex of the windmill graph on n vertices, $\mathrm{Wd}\,(3,(n-1)/2)$, and the maximum for n even is attained for the maximum degree vertex of the graph obtained from the windmill graph $\mathrm{Wd}\,(3,(n-2)/2)$ on $n-1$ vertices by subdividing an edge joining two vertices of degree 2 by a new vertex of degree 2.



(a) (b)

Figure 1: Extremal graphs within the family $\mathcal{G}_2$:
(a) for n odd and (b) for n even.

For the purposes of the following theorem, we define a new function as follows: let θ be a function such that, for any integer n,

$$\theta(n) = \begin{cases} 5/2 & \text{if } n \equiv 0 \pmod 3, \\ 0 & \text{if } n \equiv 1 \pmod 3, \\ 4/3 & \text{if } n \equiv 2 \pmod 3. \end{cases} \tag{4}$$

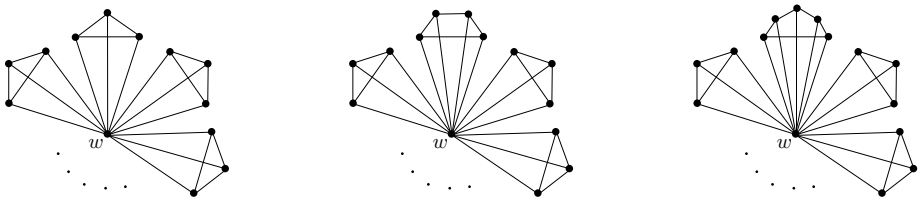
Then the following result holds.

Theorem 5. *(·, Hurajová, Madaras, Škrekovski)*

If G is a graph on n vertices with minimum degree at least 3, then

$$b_{\max}(G) \leq \frac{(n-1)(n-4)}{2} - \theta(n). \tag{5}$$

The maximum for $n \equiv 1 \pmod 3$ is attained for the central vertex of the windmill graph $\mathrm{Wd}\,(4,(n-1)/3)$, the maximum for $n \equiv 2 \pmod 3$ and $n \equiv 0 \pmod 3$ is attained for the maximum degree vertex of a graph which is obtained from the windmill graph by replacing one copy of the graph K_4 by the graph of 4- and 5-sided pyramid, respectively.



(a) (b) (c)

Figure 2: Extremal graphs within the family $\mathcal{G}_3$:
(a) $n \equiv 1 \pmod 3$, (b) $n \equiv 2 \pmod 3$ and (c) $n \equiv 0 \pmod 3$.

The investigation of values of the betweenness centrality within the family of graphs with prescribed diameter can be of great significance for the real-world networks that exhibit the small-world phenomenon. We say that a network exhibits the small-world phenomenon if, roughly speaking, any two objects (resp. individuals) in the network are

likely to be connected through a short sequence of intermediate objects (resp. individuals). Seen from a graph-theoretical perspective, the considered graph exhibits small-world phenomenon if it has a small diameter.

Hence, next we consider the family $\mathcal{D}_D$ of all graphs with diameter D. Since $\mathcal{D}_1$ consists of complete graphs and $\mathcal{D}_2$ contains all stars $K_{1,n-1}$, we have $B_{\max}(\mathcal{D}_1, n) = 0$ and $B_{\max}(\mathcal{D}_2, n) = \binom{n-1}{2}$. We now consider the general case, i.e. the family $\mathcal{D}_D$ for $D \geq 3$.

Theorem 6. *(·, Hurajová, Madaras, Škrekovski)*

Let D be an integer such that $D \geq 3$ and G be a graph with diameter D. Furthermore, let η be a function such that, for any integer k, $\eta(2k) = 0$ and $\eta(2k+1) = -1/4$. Then,

$$b_{\max}(G) \leq \frac{(n-1)(n-2)}{2} - \frac{D(D-2)}{4} + \eta(D).$$

Moreover, the maximum value is obtained only at the vertex with the largest degree, within the graph obtained by identifying a central vertex of the path P_{D+1} with the central vertex of the star graph $K_{1,n-D-1}$.

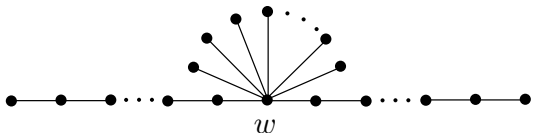


Figure 3: Extremal graph within the family $\mathcal{D}_D$, where $D \geq 3$.

4 Computational tests

In this section we describe some illustrative computations with real-world networks, focusing on the main topological properties and betweenness centrality.

For the purpose of the computational tests that follow, we choose several real-world networks that are commonly used by other researchers. These networks include: Zachary's karate club [15]; Dolphins [11]; Political books; Network science [12]; Yeast [10]; US power grid [14].

Network G	Number of vertices, n	Number of edges, m	Minimum degree, $\delta(G)$	Maximum degree, $\Delta(G)$	Diameter D	Connectivity $\kappa(G)$
Karate club	34	78	1	17	5	1
Dolphins	62	159	1	12	8	1
Political Books	105	441	2	25	7	2
Network Science	379	914	1	34	17	1
Yeast	1458	1948	1	56	19	1
US Power Grid	4941	6594	1	19	46	1

Table 1: The main topological properties of the considered real-world networks.

The definition of graph used in this paper (see Section 2) does not allow directed edges, multiple edges or self-loops. For this reason, each network is transformed so that it satisfies previously mentioned properties. In Table 1 we give an overview of used networks, together with the main topological properties. The sizes of the considered networks range from 34 vertices up to nearly 5000 vertices, and from about 78 edges up to more than 6500 edges. Moreover, all of the networks are sparse – their densities range from about 0.001 for the US power grid network, up to nearly 0.14 for the Karate club network. Each of the considered real-world network, with the exception of the third one, has minimum degree 1. Since $\kappa(G) \leq \delta(G)$, this is also the reason why each network, with the exception of the third one, has connectivity equal 1.

In order to compare the properties of real-world networks with the results obtained in the previous section, we calculate the minimum, average, median and maximum betweenness centrality for each of the considered network. Overview of the results is shown in Table 2.

Network G	Minimum $b'_{\min}(G)$	Average $\overline{b}'(G)$	Median $b'_{Me}(G)$	Maximum $b'_{\max}(G)$
Karate club	0	0.0440	0.0026	0.4376
Dolphins	0	0.0393	0.0216	0.2482
Political Books	0	0.0202	0.0044	0.1395
Network Science	0	0.0134	0.0000	0.3972
Yeast	0	0.0040	0.0000	0.2130
US Power Grid	0	0.0036	0.0004	0.2884

Table 2: Minimum, average, median and maximum betweenness centralities.

It should be noted that absolute betweenness centrality is correlated with the size of the network. A central vertex in a small network would have a smaller value of absolute centrality measure than a peripheral vertex in a large network. To allow comparison across different networks, it is important to normalize these measures by the size of the network. Therefore, the calculated centrality measures $b'(G)$ are relative, i.e. real numbers between 0 and 1. They are obtained by dividing absolute value $b(G)$ with the greatest possible value $\binom{n-1}{2}$ – the betweenness centrality of central vertex of the star, $K_{1,n-1}$.

Focusing our attention on Table 2, it can be observed that each of the considered real-world network has the relative minimal betweenness $b'_{\min}(G)$ approaching zero. Thus, the absolute $b_{\min}(G)$ is very small compared to $\binom{n-1}{2}$. The largest $b'_{\max}(G)$ is obtained within the smallest real-world network, the Zachary Karate Club network, while the lowest one is obtained in the US power grid sub-network. Obviously, the relative betweenness systematically decreases with size, since it is obtained as the normalization of the absolute value by the size of the network. The average betweenness is small compared to the maximum value, indicating that the large proportion of vertices have very small relative betweenness centrality. The same conclusion can be derived from the median values of the betweenness centrality.

Table 3 presents calculated values of the relative upper bounds $B'_{\max}(\mathcal{H}, n)$ for all fami-

Network G	$B'_{\max}(\mathcal{G}^{\Delta}, n)$	$B'_{\max}(\mathcal{C}_2, n)$	$B'_{\max}(\mathcal{C}_3, n)$	$B'_{\max}(\mathcal{G}_2, n)$	$B'_{\max}(\mathcal{G}_3, n)$	$B'_{\max}(\mathcal{D}_D, n)$
Karate club	0.971	0.910	0.906	0.968	0.900	0.992
Dolphins	0.932	0.951	0.950	0.980	0.951	0.993
Political Books	0.969	0.971	0.971	0.990	0.980	0.998
Network Science	0.973	0.992	0.992	0.997	0.992	0.999
Yeast	0.983	0.998	0.998	0.998	0.986	1.000
US Power Grid	0.948	0.999	0.999	1.000	0.991	1.000

Table 3: The maximum values of the betweenness centrality within the appropriate families.

lies of graphs $\mathcal{H}$ that include the observed real-world networks. Relative values are obtained by dividing absolute value $B_{\max}(\mathcal{H}, n)$ with $\binom{n-1}{2}$. Moreover, the upper bound for $\mathcal{C}_2$ (resp. $\mathcal{C}_3$) is obtained from the 2-core (resp. 3-core) subnetwork of the original network, where k-core subnetwork represents a maximal subnetwork in which each vertex has degree at least k. As can be seen in Table 3, all upper bounds are very high compared to the maximum values of relative betweenness obtained within the real-world networks. Since networks in the real world show structures that are far from the special structures of the extremal graphs described in Section 3, the obtained results are expected.

Nevertheless, maximum values of the betweenness centrality measure can be very useful for the procedure of normalization within different graph families. The primary purpose of normalization is to enable comparison of centrality values for individual vertices in different networks. Moreover, the obtained maximum values may be used to calculate Freeman's general measure of centrality [6], which indicates the centralization of a network G as the absolute deviation from the maximum betweenness value on G. This measure is simply the sum of the differences in centrality between the most central vertex and all the others, normalized by the maximum possible value over all connected graphs from the corresponding family.

References

[1] J. A. Bondy, U. S. R. Murty, *Graph Theory*, Graduate Texts in Mathematics 244, Springer, 2008.

[2] F. Comellas, S. Gago, *Spectral bounds for the betweenness of a graph*, Linear Algebra Appl. **423** (2007), 74–80.

[3] M. G. Everett, P. Sinclair, P. A. Dankelmann, *Some centrality results new and old*, J. Math. Sociol. **28**(4) (2004), 215–227.

[4] M. G. Everett, S. P. Borgatti, *Ego network betweenness*, Social Networks **27** (2005), 31–38.

[5] L. C. Freeman, *A set of measures of centrality based on betweenness*, Sociometry **40** (1977), 35–41.

[6] L. C. Freeman, *Centrality in networks: I. Conceptual clarification*, Social Networks **1** (1979), 215–239.

[7] S. Gago, *Métodos espectrales y nuevas medidas modelos y parámetros en grafos pequeño-mundo invariantes de escala*, Ph.D. Thesis, Universitat Politècnica de Catalunya, 2006 (in Spanish).

[8] S. Gago, J. Hurajová, T. Madaras, *Notes on the betweenness centrality of a graph*, Math. Slovaca **62** (1) (2012), 1–12.

[9] S. Gago, J. Hurajová, T. Madaras, *On betweenness-uniform graphs*, Czech. Math. J. (2013)

[10] H. Jeong, S. P. Mason, A. L. Barabsi, Z. N. Oltvai, *Lethality and centrality of protein networks*, Nature **411** (2001), 41–42.

[11] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, *The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations*, Behav. Ecol. Sociobiol. **54** (2003), 396–405.

[12] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Phys. Rev. E **74** (2006), 036104, 2006.

[13] P. Sinclair, *Betweenness centralization for bipartite graphs*, J. Math. Sociol. **29** (2005), 25–31.

[14] D. J. Watts, S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature **393** (1998), 440–442.

[15] W. W. Zachary, *An information flow model for conflict and fission in small groups*, J. Anthropol. Res. **33** (1977), 452–473.

Diameter on some classes of fullerene graphs

Vesna Andova
Faculty of Electrical Engineering and Information Technology
Ss Cyril and Methodius Univ.
Ruger Boskovik bb, Skopje, Macedonia
vesna.andova@gmail.com
František Kardoš
LaBRI
University of Bordeaux
351, cours de la Libération, 33405 Talence, France
frantisek.kardos@labri.fr
Riste Škrekovski
Faculty of Information Studies
Talcev 3, Novo mesto, Slovenia
skrekovski@gmail.com

Abstract. *Fullerene graphs are 3-connected cubic planar graphs with only pentagonal and hexagonal faces. This paper is a collection of results concerning the diameter of some classes of fullerene graphs.*

Keywords. fullerene graphs, nanotubes, icosahedar, diameter

1 Introduction

Fullerene graphs are 3-connected, cubic, planar graphs with only pentagonal and hexagonal faces. From Euler formula follows that this graph have precisely 12 pentagons. and α hexagonal faces, $\alpha \neq 1$. Grünbaum and Motzkin [11] showed that fullerene graphs on n vertices exist for all even $n \geq 24$ and for $n = 20$. The smallest fullerene graph, the dodecahedron, is a graph on 20 vertices comprised only by pentagonal faces.

Although the number of pentagonal faces is negligible compared to the number of hexagonal faces, their position is crucial for the shape of a fullerene graph. If the pentagons are grouped into two rather compact patches, we obtain a class of fullerene graphs of tubular shapes, called *nanotubes*. On the other hand, if the pentagons are distributed such that their centers are vertices of a regular icosahedra, we obtain fullerene graphs of spherical shape with icosahedral symmetry, whose smallest representative is the dodecahedron.

The interest for studying fullerene graphs started after the discovery of the first fullerene molecule C_{60} in 1985 by Kroto et al. [14]. The first fullerene molecule is also known as buckminsterfullerene, named after Richard Buckminster Fuller, whose geodetic domes it resembles. Due to the wide spectrum of possible applications, fullerenes attract the at-

tention of diverse research communities. The main aim of their studies is to determine the physical and chemical properties of a particular fullerene. As each molecule can be represented as graph, some graph invariant can predict different properties of the compound, what makes them interesting from graph theory perspective [1, 6, 8, 9, 12].

Here we pay our attention of to the diameter of a fullerene graph. The results presented here are overview of the results from [2, 3, 4, 5].

The *distance* between two vertices $u, v \in V(G)$ in a connected graph G is the length of the shortest path between u and v, and it is denoted by $d(u, v)$. The *diameter* of a connected graph G, $\mathrm{diam}(G)$, is the maximum distance between two vertices of the graph G.

2 Some classes of fullerene graphs

Even though the number of pentagons is strictly determined, while there is no limitation on the number of hexagons, the fullerene is shaped by the position of the 12 pentagons. According to the position of the pentagons, we distinguish three classes of fullerene graphs: nanotubical, icosahedral, and isolated pentagon fullerene graphs.

2.1 Icosahedral fullerene graphs

The common feature of all icosahedral fullerenes is their spherical shape. Goldberg [10] observed that all icosahedral fullerene graphes can be obtained by mapping (a part of) the hexagonal grid onto the triangular faces of an icosahedron, and showed that the number of vertices n in a polyhedron of icosahedral symmetry can be determined by two integers i and j by the following equation, conveniently called the *Goldberg equation*

$$n = 20(i^2 + ij + j^2). \tag{1}$$

The integers i and j in the Goldberg equation are in fact the components of a two-dimensional *Goldberg vector* $\vec{G} = (i, j)$. To avoid the mirror effect, we always assume that $0 \leq i \leq j$ and $0 < j$.

This vector determines the distance and positions of the vertices of the (i, j)-triangle in the hexagonal lattice. Precisely 20 such (i, j)-triangles produce an (i, j)-icosahedral fullerene in a manner shown on Figure 1. The vertices of the triangles are centers of the 12 pentagons of the fullerene graph.

The *icosahedral group* $\mathbf{I}$ is the group of rotational symmetries of the icosahedron and dodecahedron, of order 60. This group is isomorphic to $\mathbf{A}_5$, the alternating group of even permutations of five objects. The *full icosahedral group* $\mathbf{I}_h$, is the point group of all symmetries of the icosahedron and dodecahedron. This group is equivalent to the group direct product $\mathbf{A}_5 \times \mathbb{Z}_2$ of the alternating group $\mathbf{A}_5$ and cyclic group $\mathbb{Z}_2$. The order of the full icosahedral group of symmetries is 120. The (i, i)- and $(0, i)$-icosahedral fullerene graphs, $i > 0$, have full icosahedral symmetry group, i.e., every element of this class of graphs has a symmetry group $\mathbf{I}_h$.

2.2 Nanotubical fullerene graphs

While the icosahedral fullerenes have “spherical" shape, there is a class of fullerene graphs of tubular shapes, called *nanotubical* graphs or simply *nanotubes*. They are cylindrical in shape, with the two ends capped with a subgraph containing six pentagons and

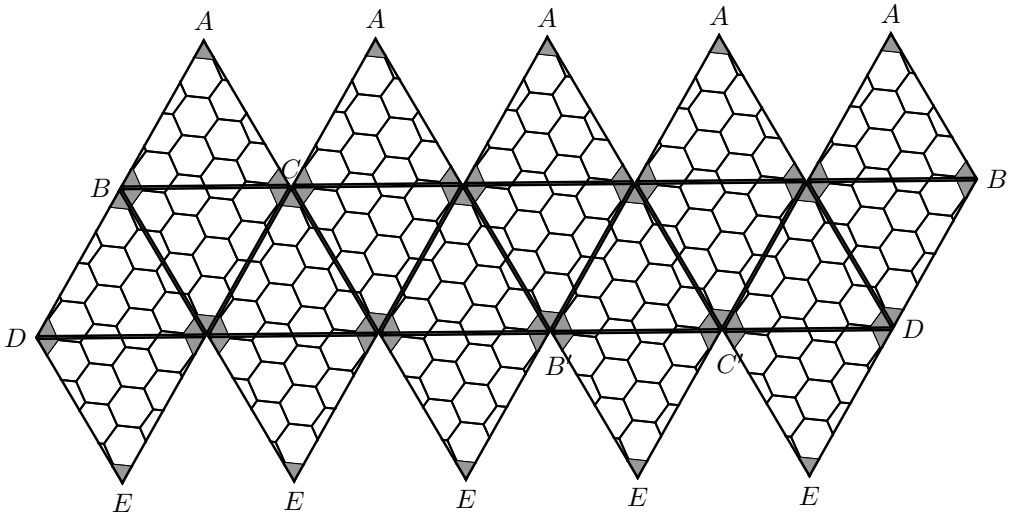


Figure 1: A $(2,3)$-icosahedral fullerene. The vertices with a same label coincide. The vertices of each triangular are centers of the pentagons.

possibly some hexagons. The cylindrical part of the nanotube can be obtained by rolling a planar hexagonal grid. The way the grid is wrapped is represented by a Goldberg vector (i,j), called also the *type* of the nanotube. Nanotube fullerene graphs exist for all vectors (i,j) with $i+j \geq 6$ and for the vector $(0,5)$. The the sum $p+q$ is called *circumference* of a (p,q)-(nano)tube.

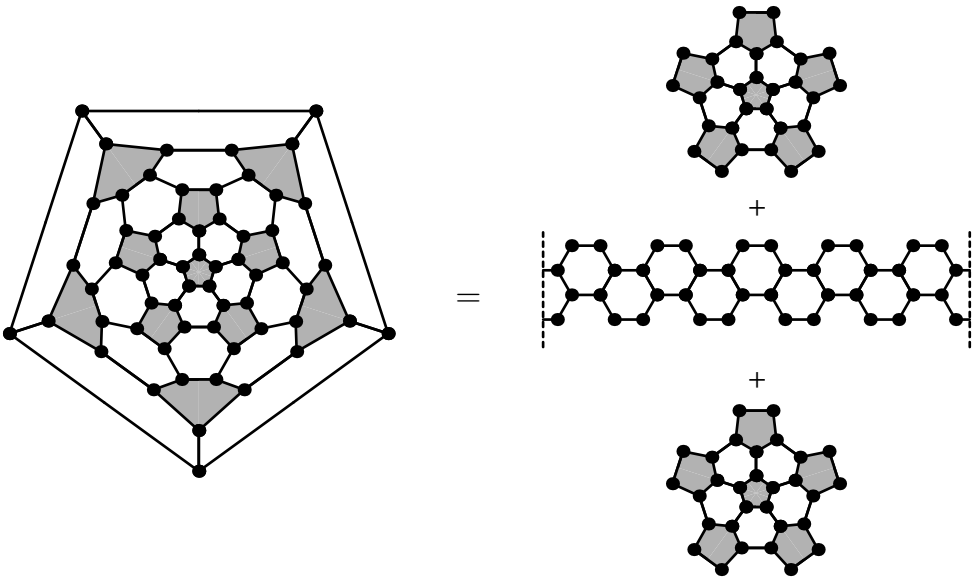


Figure 2: Buckminsterfullerene is the smallest $(5,5)$-nanotube.

2.3 Isolated pentagon fullerene graphs

There are fullerene graphs where no two pentagons are adjacent, i.e. each pentagon is surrounded by five hexagons. Those fullerene graphs satisfy the *isolated pentagon rule* or shortly IPR, and they are the most stable fullerene compounds [13]. The Buckminsterfullerene C_{60} can be viewed as the smallest $(5,5)$-nanotube, see Figure 2. If the caps from

C_{60} are chosen differently, as shown on Figure 3, this fullerene can also be considered as $(9,0)$-nanotube. In both cases it is the smallest nanotube with the caps satisfying the IPR. There are such nanotubes for all types (i,j) with $i+j \geq 11$ and for $(5,5)$, $(0,9)$, $(0,10)$, $(1,9)$, and $(2,8)$.

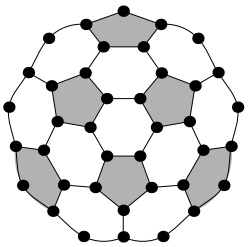


Figure 3: Cap of a $(9,0)$-isolated pentagon nanotube.

3 Diameter of fullerene graphs

For large enough fullerene graphs of spherical shape, the diameter $\mathrm{diam}(G)$ is proportional to the radius of the sphere, whereas the number of vertices n is proportional to its surface. Hence, one could expect the diameter of such graphs to be of order $\Theta(\sqrt{n})$.

The bounds of the diameter of general fullerene graphs was solved in [3], and the following theorem confirms the expected results.

Theorem 1 [3] *Let G be a fullerene graph with n vertices. Then,*

$$\mathrm{diam}(G) \geq \frac{\sqrt{24n-15}-3}{6}\,.$$

The diameter of the special case of highly symmetric icosahedral fullerenes was settled in [5].

Theorem 2 [5] *Let G be a $(0,i)$-icosahedral fullerene graph with $i>0$. Then* $\mathrm{diam}(G) = 6i-1 = \sqrt{\frac{9}{5}n}-1\,.$

Theorem 3 [5] *Let G be an (i,i)-icosahedral fullerene graph with $i>0$. Then* $\mathrm{diam}(G) = 10i-1 = \sqrt{\frac{5}{3}n}-1.$

As we believe that the icosahedral fullerenes have the smallest diameter, it leads us to conjecture the following:

Conjecture 1 *For every fullerene graph F on n vertices holds* $\mathrm{diam}(F) \geq \left\lfloor \sqrt{\frac{5}{3}n} \right\rfloor - 1.$

The upper bound of the diameter is general fullerene graph is determined in [3].

Theorem 4 [3] *Let G be a fullerene graph with n vertices. Then,*

$$\mathrm{diam}(G) \leq \frac{n}{5}+1\,.$$

Observe that the upper bound of the diameter is obtained for a $(5,0)$-nanotube on 20 vertices, the dodecahedron. For a nanotubical fullerene graph of type (i,j), the diameter $\mathrm{diam}(G)$ is proportional to the length of the cylindrical part of the graph, whereas the number of vertices n is proportional to the product of the length of the tube and its circumference $i+j$. In this case, one could expect the diameter of such graphs to be of order $\Theta(n)$ [2].

Theorem 5 [2] *Let F be a (p,q)-nanotube, $p,q \in \mathbb{N}$, on n vertices. Then, there are constants $C'_{p,q}$ and $C''_{p,q}$ such that*

$$\frac{n}{p+q} + C'_{p,q} \leq \mathrm{diam}(F) \leq \frac{n}{p+q} + C''_{p,q}\,.$$

Next we determine the diameter of isolated pentagon $(9,0)$-nanotubes as IPR nanotubes with the smallest circumference [2]. Observe that the Buckminsterfullerene can be considered as an isolated pentagon $(9,0)$-nanotube on 60 vertices, as nanotube with no hexagonal rings. The diameter of C_{60} is 9, i.e. it holds $\mathrm{diam}(C_{60}) = \frac{n+21}{9}$.

Theorem 6 [2] *Let F be an isolated pentagon $(9,0)$-nanotube on n vertices. Then,*

$$\frac{n+3}{9} \leq \mathrm{diam}(F) \leq \frac{n+21}{9}\,.$$

As isolated pentagon $(9,0)$-nanotube has the smallest circumference among all isolated pentagon nanotubes on n vertices, from Theorem 5 follows that among all isolated pentagon nanotubes on n vertices, $(9,0)$-nanotube has the largest diameter. This observation gives us the upper bound on the diameter for isolated pentagon nanotubes. Notice that this bound is tight only for isolated pentagon fullerenes on 60 and 78 vertices [7].

Corollary 1 *Let F be an IPR nanotube on n vertices. Then*

$$\mathrm{diam}(F) \leq \frac{n+21}{9}\,.$$

4 Acknowledgments

Research supported in part by ARRS Program P1-00383, France-Slovenian grant BI-FR/14-15-PROTEUS-001, and Creative Core - FISNM - 3330-13-500033.

References

[1] E. Albertazzi, C. Domene, P. W. Fowler, T. Heine, C. Van Alsenoy, F. Zerbetto, Pentagon adjacency as a determinant of fullerene stability, *Phys. Chem. Chem. Phys.* 1: 2913–2918, 1999.

[2] V. Andova, D. Blankuš, T. Došlić, R. Škrekovski, On diameter of nanotubical fullerene graphs, to appear in *MATCH Commun. Math. Comput. Chem.*

[3] V. Andova, T. Došlić, M. Krnc, B. Lužar, R. Škrekovski, On the diameter and some related invariants of fullerene graphs, *MATCH Commun. Math. Comput. Chem.* 68:109–130, 2012.

[4] V. Andova, F. Kardos, R. Škrekovski, Fullerene Graphs and Some Relevant Graph Invariants, In: I. Gutman (Ed.), *Topics in Chemical Graph Theory*, Kragujevac, 2014.

[5] V. Andova, R. Škrekovski, Diameter of full icosahedral-symmetry fullerene graphs, *MATCH Commun. Math. Comput. Chem.* 70:205–220, 2013.

[6] T. Došlić, Bipartivity of fullerene graphs and fullerene stability, *Chem. Phys. Lett.* 412:336–340, 2005.

[7] T. Došlić, Empirical Study of Diameters of Fullerene Graphs, In:Mihai V. Putz (Ed.), *Carbon Bonding and Structures, Carbon Materials: Chemistry and Physics, Springer Netherlands* (2011).

[8] T. Došlić, T. Reti, Spectral properties of fullerene graphs, *MATCH Commun. Math. Comput. Chem.* 66:733–742, 2011.

[9] T. Došlić, D. Vukičević, Computing the bipartite edge frustration of fullerene graphs, *Discrete Appl. Math.* 155:1294–1301, 2007.

[10] M. Goldberg, A class of multi-symmetric polyhedra, *Tohoku Math. J.* 43:104–108, 1939.

[11] B. Grünbaum, T. S. Motzkin, The number of hexagons and the simplicity of geodesicson certain polyhedra, *Can. J. Math.* 15:744–751, 1963.

[12] S. Fajtlowicz, C. E. Larson, Graph-Theoretic Independence as a Predictor of Fullerene Stability, *Chem. Phys. Letters* 377:485–490, 2003.

[13] H. W. Kroto, The stability of the fullerenes C_n, with $n = 24, 28, 32, 36, 50, 60$ and 70, *Nature* 329:529–531, 1987.

[14] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, R. E. Smalley, C60: Buckminsterfullerene, *Nature* 318:162–163, 1985.

Comparison of SAS/STAT procedure and Variable Neighbourhood Search based clustering applied on Telecom Serbia data

Stefana Janicijevic, Dragan Urosevic, Nenad Mladenovic
Faculty of Technical Sciences
University of Novi Sad
Trg Dositeja Obradovica, 21000 Novi Sad, Serbia
{stefana, draganu, nenad}@mi.sanu.ac.rs

Abstract: *In this study we analyze possibility of improving decision qualities in largest Serbian telecommunication provider, Telecom Serbia a.d. For that purposes we compare qualities of SAS/STAT clustering with Variable Neighborhood Search based methods on three months real data set of prepaid customers. It appears that Variable Neighborhood Search based approach provides results with smaller error, giving more precise decision support to marketing department.*

Key Words: *Variable Neighborhood Search, Clustering, SAS/STAT procedures, Decision Support*

1 Introduction

We propose paper according to research which was connected with Telekom Srbija a.d. data base, with an aim to using advanced technologies of business analysis based to coding, programing, statistics, researching and mathematical modeling of data.
The assignment was detection of pattern and information relevant for decision makers, since those kind of rules in data could be used in business strategies and marketing visions further.
After detailed analysis of the sample of particular criterion we propose the models which minimize CPU time and efforts and maximize the effects.
The research was expanded and resulted to several directions activity, because we recognize many challenges of mathematical programing and optimization.
The business models are necessary in the world of leading corporations, the business prestige competition and profit creating, according to original and clear ideas. Good practice shows that decisions in huge companies have to be based on applied mathematics, automated processes and data mining procedures as it could be precisely found relevant information in the heap of recorded data, so information and judgments could be measured clearly.
In the next sections we are giving brief description of analyzed data and present the models which are developed with intention that strategically mechanisms could be recognized and scaled, so we could efficiently use the latest scientific achievements in mathematical modeling and apply them in industry.

Telekom Srbija a.d. company has plenty possibilities to develop business intelligence tools and to develop the predictive models for various fields of marketing requests.
The assignment of customer profiling approach was conducted in several manners such as:

- Statistically: we start from descriptive statistical models which are based on simply explanations of the main parameters (univariate analysis), through more complex algorithmic methods, tests and predefined procedures which are available above SAS servers and libraries implemented to SAS programing language
- Algorithmically: with no statistical tools, but only metaheuristic methods of solution space searching implemented in various programing languages

After that, according to looking at prepaid data base of customers we build up the idea of grouping and classifying customers by similar criterions in order that we would manage foreseeable and easier customer segmentation which is included in strategic plan of particular areas like campaigns, service growth and facilities in company Telekom Srbija a.d.
It is necessary to move from the point of descriptive statistic to advanced modeling point through using modern big data, data mining and data wrangling techniques or just using heuristic algorithms in solution search space as it is inspired of customer offer personalization and profiling every single customer or creating personal customer card.
Research in this paper was applied in two directions. First one was cluster forming via SAS/STAT packages (see SAS Campus 2008), official and depute software SAS Enterprise Guide 5.1which is available in company Telekom Srbija a.d.
We program in SAS Base program language with running statistical library procedures for cluster analysis for sample of the specific population.
Second direction was metaheuristic Variable Neighborhood Search (VNS) application. We formulated problem as p-median searching for complete graph. Graph was constructed as average of three months for two criterions. One criterion was recharge revenue and second was number of days between every recharge.
This formulation was solved with VNS method in C++/gcc program language.
Comparative analysis appointed that VNS method assigned results with smaller error in respect to SAS procedures of hierarchical clustering.
Input data, as we said, for both approaches were identical and based on three months average of revenue recharge and three months average of number of days between recharge.
Conception of this paper is like that in next section we are explaining briefly clustering techniques supported with SAS/STAT package which is professionally used in company. In the third section there is described the way of problem formulation as p-median problem and the VNS method for solving.

2 SAS/STAT clustering procedures

SAS/STAT library completed and predefined procedures. Every generalization of cluster analysis in SAS package should be tested since there are manifold methods of clustering for various industries which are developed in SAS Institute (see SAS Campus 2008). There are defined particular cluster formulations and particular cluster functions depend on cluster hierarchy. Diversification of clustering techniques SAS Institute fitted with classification, clamping, machine learning, measurement of shapes, partitioning, systematic, typology, pattern recognition, vector quantization, etc.

SAS improved several types for cluster methods as disjoint clusters and several types of hierarchical clusters for the necessity of business modeling.
Beside these there are another types of clusters (fuzzy, overlapping, etc.). SAS predefined procedures process the data through several forms: coordinate matrix, L_2 data distances, correlation matrix, covariance matrix, etc. The most important procedures which are used in SAS cluster library: Cluster, Fastclus, Modeclus, Varclus, and Tree. In the framework of SAS/STAT procedures there are possibilities to create another types of cluster methods, though, at the aspect of business logic, they are not simple, functional or efficient in using. It is necessarily to invest in data preparation and data quality, as in result evaluation. For example, PROC FASTCLUS should be used for every number of cluster recounts. PROC CLUSTER is not efficient enough for large number of data (for data base of terabyte dimensions as in Telekom Srbija a.d. company).

2.1 Centroid method of SAS/STAT library

Centroid method is applied firstly on 5 clusters, and then on 10 clusters. Results of clustering procedures against data base of 23.111 records could be analyzed in the Table 1.1 CPU time needs for running exe program could be analyzed in the Table 1.2.
Let x_i is one observation. D_{KL} is anyone distance or various measure between clusters C_K i C_L. N_K is number of observations in cluster C_K, while N_L is number observations in cluster C_L. Then centroid method could be formulated as (see SAS Campus (2008)):

$$D_{KL} = \left\| x_{K-} x_L \right\|^2$$

$$D_{JM} == \frac{N_K D_{JK} + N_L D_{JL}}{N_M} - \frac{N_K N_L D_{KL}}{N_M^2}$$

Where is

$$d(x, y) = \| x - y \|^2$$

distance of two centroids.

Among this method, we tested another methods which are defined in mentioned library, such as K-MEAN, WARD, AVERAGE, SINGLE i DENSITY. Testing resulted that the complexity parameter CPU time is about 43 minutes.

3 Variable Neighbourhood Search method for the p - median

Variable neighborhood search method is famous metaheuristic which was proposed in 1995. at the first version, and after that method was implemented in many variations, thanks to successfully applications through the world (see Mladenovic 1997). The construction of the basic method consists of the systematically searching of the solution space. It is necessary to define metrics, or distance between solutions. After that, we define finite set of neighborhoods N_k , $k = 1,\ldots,k_{\max}$. If it is $x \in X$ arbitrary

solution, then $N_k(x) \subset X$ is the set of solutions which exist in k-th neighborhood of solution x. Neighborhood $N_k(x)$ is consisted from all solutions which are with the k distance from the solution x (see Hansen (2008)).
Within the concrete assignment of clustering, we applied already implemented method of the variable neighborhood search for p - median problem (see Mladenovic (2007)). The customers are treated as two-dimensional points, as we managed with two numeric variables. But, in general sense, if we could have created k variables, that would be the points in a k - dimensional space.
The distance is calculated as Euclidean distance in two-dimensional space. Problem is formulated as selection of p points (or p centers) from this set, so the sum of the distances from every residual point to the nearest center is minimum. VNS method implies the local search and shaking, so we implemented both phase in the p - median algorithm.
Local search consists from swap transformation application, or exchange one center with another point (not center point), until performing improvement is made.
During the search we are using strategy of the first improvement which means, if an improvement is obtained, k is returned to its initial value and the new incumbent updated.
Shaking in the k neighborhood is consists from k applications interchange existent center with some point not-center.
Testing resulted several parameters of complexity: CPU time about 30 minutes and maximum neighborhood shaking is $k_{\max} = \frac{p}{2}$ (p is number of clusters).

4 Results

As it could be seen, SAS/STAT procedure R-Square is 21.6% for 5-cluster method, and for VNS method it is 53% for the same parameter, what creates conclusion that the SAS model less managed goodness of fit than VNS model. CCC parameter is far from the positive values what indicates that goodness of fit is not well.
For the SAS/STAT procedure for 10-cluster method R-Square is 77.4%, comparing with VNS method, it is the same value. CCC parameter is again far from the positive values and that indicates that goodness of fit is not well.
CPU time performing was better (faster) for the VNS method comparing with SAS method.
SAS procedures for clustering present that their models should be fitted and repaired, since, model proposes correct result only for the more than 30 number of clusters. That is really huge number of clusters since marketing of the company should propose more than 30 campaigns, so that is expensive for the management.
Data analyst should make interventions on the data and create artificially records to make well result with this kind of package.
We did also one sort of artificially model, as we clustering the first cluster recursively, since this cluster is the most considerable. In that way, model gave satisfying results, but since it is spurious made and coded with the aim, we cannot speak about example for this kind of data, so we cannot propose directly using
SAS package is not enough for well results and it is not perfect for the programming self-functions.

Table 1: Parameters of model

	SAS/STAT Centroid-5 CLUSTER	SAS/STAT Centroid-10 CLUSTER	VNS P-median 5	VNS P-median 10
R-square	0.216	0.774	0.534	0.77
Semipartial R- square	0.0014	0.0001		
Cubic Clust Crit	-271	-141		
Norm Centroid Distance	3.2802	1.2863		

Table 2: Customer number in cluster 5

Cluster	SAS Customer number in cluster	VNS Customer number in cluster
1	23054	4944
2	46	833
3	9	1879
4	1	10763
5	1	4692

Table 3: Customer number in cluster 10

Cluster	SAS Customer number in cluster	VNS Customer number in cluster
1	21678	3862
2	833	2614
3	543	2405
4	37	227
5	4	1107
6	9	665
7	3	3572
8	2	833
9	1	1605
10	1	6221

5 Conclusion

The aim of the optimization and modeling with cluster method is in clever algorithms, if data features (points, records) are coded with automatic searching in optimal time and

memory using and with so homogenous points inside groups and heterogeneous points between groups.
If we need fuzzy clusters as we could position particular points with some probabilities in several various groups, we should propose model which could independently scores probability of points repositioning in groups. Well-developed algorithm minimizes prejudice or statistical bias. Statistical bias is a kind of deviation since it is calculated that it is systematically distinct from population constant.
This implies that the model is statistically and analytically well, if there are low capacity of preconceived defined parameters and also if it is developed without obligation for result improvement or data quality.
Good model considers mapping of acnodes which are significant and independent, so result could appear by using properly features of the data base significant criterions, without outside manual interventions by analyst.
Alongside, officially accepted parameters, as R-square, consider that analyst should determine more stable model by comparing of them.

6 References

[1] Hansen, P.; Mladenovic, N.; Perez, J. M. Variable neighbourhood search: methods and applications. 4OR, 6, 319-360, 2008.
[2] Mladenovic, N.; Hansen, P. Variable Neighborhood Search. Computers OR, 24, 1097-1100, 1997.
[3] Mladenovic, N.; Brimberg, J.; Hansen, P.; Perez, J.M. The p-median problem: A survey of metaheuristic approaches. European Journal of Operational Research, 179, 927-939, 2007.
[4] SAS Campus. SAS/STAT 9.2 User 's Guide. (1st ed.). SAS Institute Inc, 2008.

On parity and weak-parity edge-colorings

Borut Lužar
Faculty of Information Studies
Ulica talcev 3, Novo mesto, Slovenia
borut.luzar@gmail.com
Mirko Petruševski
Department of Mathematics and Informatics, Faculty of Mechanical Engineering
Ss Cyril and Methodius Univ.
Ruger Boskovik bb, Skopje, Macedonia
mirko.petrushevski@gmail.com
Riste Škrekovski
Department of Mathematics
University of Ljubljana, Slovenia
Faculty of Information Studies
Ulica talcev 3, Novo mesto, Slovenia
Faculty of Mathematics, Natural Sciences and Information Technologies
University of Primorska, Slovenia
skrekovski@gmail.com

Abstract. *A graph is odd if each of its vertices is of odd degree. An odd edge-coloring of a graph G is a (not necessarily proper) edge-coloring such that each color class induces an odd subgraph of G. Any mapping $\pi : V(G) \to \{0, 1\}$ is a vertex signature of G. An edge-coloring of G is said to be a vertex-parity edge-coloring of (G, π) if for every $v \in V(G)$ each color appearing at v appears in parity accordance with π, i.e. it appears even (resp. odd) number of times at v whenever $\pi(v)$ equals 0 (resp. 1). Similarly, a weak vertex-parity edge-coloring of (G, π) requires that at each non-isolated vertex v at least one color appears in parity accordance with π. This paper is a collection of results concerning odd edge-colorings of graphs, as well as vertex-parity and weak vertex-parity edge-colorings of graphs with prescribed vertex signatures.*

Keywords. odd edge-coloring, odd chromatic index, vertex signature, vertex-parity edge-coloring, weak vertex-parity edge-coloring

1 Introduction

Throughout the article, we consider finite graphs $G = (V(G), E(G))$ (i.e. graphs having finite sets $V(G)$ and $E(G)$ of vertices and edges, respectively) with loops and parallel edges allowed. For any pair of (distinct) adjacent vertices v, w in G, the edge set of all vw-edges is called a *bouquet* of G (more verbosely, the vw-bouquet). The maximum size

of any bouquet is the *multiplicity* of G. Whenever G is loopless and of multiplicity 1, we say it is a *simple graph*. The parameters $n(G) = |V(G)|$ and $m(G) = |E(G)|$ are *order* and *size* of G, respectively. For a graph G, each vertex v having an even (resp. odd) degree $d_G(v)$ is an *even* (resp. *odd*) *vertex* of G. A graph is *even* (resp. *odd*) whenever all its vertices are even (resp. odd). An arbitrary mapping $\varphi : E(G) \to S$ is called an *edge-coloring of* G, and then S is the *color set* of φ. For each $c \in S$, the *color class of* c is the subset of edges colored by c. The spanning subgraph of G with the edge set E_c is denoted by G_c. We say that the color c is *odd* (resp. *even*) *at* v when v is an odd (resp. even) vertex of G_c. The color c *appears at* v if v is not isolated in G_c. Here, it is important to mention that, similarly to the definition of $d_G(v)$, each loop at v which receives the color c contributes 2 to the count of appearances of c at v, i.e 'counts twice' in $d_{G_c}(v)$.

As introduced in [11], an *odd edge-coloring* of a graph G is a (not necessarily proper) edge-coloring such that each color class induces an odd subgraph of G. In other words, it is required that at every vertex of G, each appearing color is odd. Equivalently, an odd edge-coloring of G can be seen as an edge decomposition of G into (edge disjoint) odd subgraphs. Such decompositions represent a counterpart to decompositions into even subgraphs, which were mainly used while proving various flow problems (see e.g. [9, 12]). Historically speaking, as a topic in graph theory, decomposing into subgraphs of a particular kind started with [1].

In the next section we present an overview of the results from [4, 10] regarding the odd edge-colorability of general graphs (loops and parallel edge allowed). In the following section, we introduce the notion of vertex-parity edge-coloring which generalizes the concept of odd edge-coloring, and make an overview of the results from [5]. The final section contains the results from [6] regarding the related notion of weak vertex-parity edge-colorings.

2 Odd edge-coloring of graphs

If an odd edge-coloring of G uses at most k colors, we say it is an *odd k-edge-coloring*, and then G is *odd k-edge-colorable*. Whenever G admits odd edge-colorings, the *odd chromatic index* $\chi'_o(G)$ is defined as the minimum integer k for which G is odd k-edge-colorable. Clearly, any simple graph admits odd edge-colorings (since we could merely color each edge with a different color). Moreover, an obvious necessary and sufficient condition for the odd edge-colorability of a (not necessarily simple) graph G is the absence of vertices incident only to loops. Apart from this, the presence of loops does not influence the existence nor changes the value of $\chi'_o(G)$. Therefore, the study of odd edge-colorability may be restricted to connected loopless graphs without any loss of generality. The odd edge-colorability of simple graphs was considered in [11], and the following result was proved.

Theorem 2.1. [11] *Every simple graph G is odd 4-edge-colorable.*

The graph W_4, a wheel with 4 spokes, cannot be decomposed into less than 4 odd subgraphs. This example can be used to construct an infinite series of related examples, as shown in [8]. Hence, when the consideration is restricted to simple graphs G, the bound $\chi'_o(G) \leq 4$ is tight. However, this bound does not hold for an arbitrary connected loopless graph G. For instance, Fig. 1 depicts four graphs (belonging to the family of Shannon triangles) with the following property: each of their odd subgraphs is a copy of K_2. Hence,

the two rightmost graphs have the odd chromatic indices greater than 4.

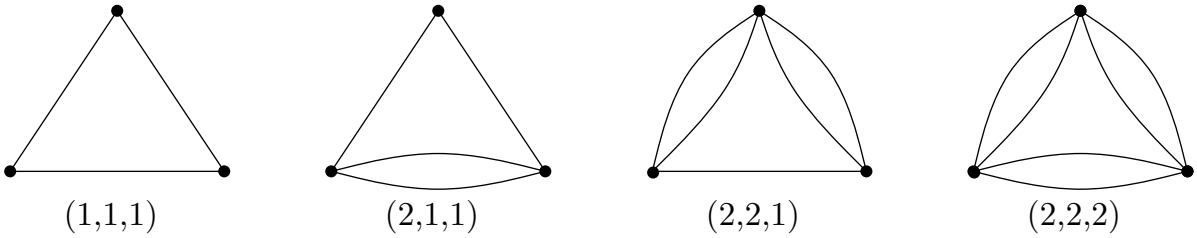


Figure 1: Examples of Shannon triangles (the smallest one of each type).

A *Shannon triangle* is a loopless graph on three pairwise adjacent vertices. The edge set of any odd subgraph of a Shannon triangle G is fully contained in a single bouquet of G. Let p, q, r be the parities of the sizes of the bouquets of G in non-increasing order, with 2 (resp. 1) denoting an even-sized (resp. odd-sized) bouquet. Then G is a Shannon triangle of *type* (p, q, r). The graphs in Fig. 1 are the smallest (in terms of size) Shannon triangles of type $(1, 1, 1)$, $(2, 1, 1)$, $(2, 2, 1)$, and $(2, 2, 2)$, respectively.

The following result from [4] states that 6 colors suffice for an odd edge-coloring of any connected loopless graph, and furthermore it characterizes when 6 colors are necessary.

Theorem 2.2. [4] *For every connected loopless graph G, it holds that $\chi'_o(G) \leq 6$. Moreover, equality is attained if and only if G is a Shannon triangle of type $(2, 2, 2)$.*

Having in mind how any odd subgraph of a Shannon triangle looks like, the following result is straightforward (see [4]).

Proposition 2.3. [4] *For any Shannon triangle G of type (p, q, r), it holds that $\chi'_o(G) = p + q + r$.*

In particular, any Shannon triangle G of type $(2, 2, 1)$ has the odd chromatic index equal to 5. The following result, coupled with Theorem 2.2, characterizes the Shannon triangles of type $(2, 2, 1)$ as the only connected loopless graphs G having $\chi'_o(G) = 5$.

Theorem 2.4. [10] *Every connected loopless graph which is not a Shannon triangle of type $(2, 2, 2)$ or $(2, 2, 1)$ is odd 4-edge-colorable.*

The last theorem clearly generalizes Theorem 2.1 to the context of general graphs. Although it is easy to state which are the connected loopless graphs G having $\chi'_o(G) = 6$ and which are having $\chi'_o(G) = 5$, it seems that a simple characterization of $\chi'_o(G) = 4$ does not exist.

3 Vertex-parity edge-colorings

Instead of requiring that at every vertex each appearing color is odd, we could designate a parity at each vertex and look for edge-colorings complying to this parity. This leads to a natural generalization of odd edge-colorings, defined (in [5]) as follows. For a graph G, a mapping $\pi : V(G) \to \{0, 1\}$ is a *vertex signature* of G, and we refer to (G, π) as a *pair*. Given a (not necessarily proper) edge-coloring φ of G and a vertex $v \in V(G)$, we say

a color c *appears* at v *in parity accordance with* π if $d_{G_c}(v)$ is greater than zero and of the same parity as $\pi(v)$. If at each non-isolated $v \in V(G)$, this holds for every appearing color, we call φ a *vertex-parity edge-coloring* of (G, π). (In the particular case when π is identically equal to 1, we speak of the *odd vertex signature* of G, and then clearly a vertex-parity edge-coloring of (G, π) is exactly the same thing as an odd edge-coloring of G.)

A vertex-parity edge-coloring of (G, π) using at most k colors is referred to as a *vertex-parity k-edge-coloring*, and then (G, π) is *vertex-parity k-edge-colorable*. When (G, π) admits a vertex-parity edge-coloring, the *vertex-parity chromatic index* $\chi'_p(G, \pi)$ is defined as the minimum integer k for which (G, π) is vertex-parity k-edge-colorable. Each pair consisting of a component of G and the respective restriction of π is a *component of* (G, π). When G is connected, we call (G, π) a *connected pair*. Any vertex of G which is mapped by π to 0 (resp. 1) is named a *zero-vertex* (resp. *one-vertex*) of (G, π).

Proposition 3.1. [5] *The following two conditions are necessary for the existence of* $\chi'_p(G, \pi)$.

$(p1)$ *Each odd vertex of* G *is a one-vertex of* (G, π).

$(p2)$ *If a component of* (G, π) *has a single one-vertex* v*, then* v *is an isolated vertex.*

Whenever the conditions $(p1)$ and $(p2)$ are fulfilled, we say that π is a *proper vertex signature* of G, or equivalently we call (G, π) a *proper pair*.

In order to present our results about vertex-parity edge-colorings, we need to introduce some specialized notions from [5]. First, let us slightly enlarge the scope of the notion of Shannon triangle from the previous section by allowing the presence of arbitrarily many loops at each of the three vertices. Given a Shannon triangle G_0, let π_0 be its odd vertex signature. Consider a subdivision G_1 of G_0, and define π_1 as the vertex signature of G_1 which agrees with π_0 on $V(G_0)$ and maps every vertex of $V(G_1) \setminus V(G_0)$ to 0. Given two zero-vertices w' and w'' of a pair (G, π), identify them into a new vertex w (every possible $w'w''$-edge becomes a loop) and define a vertex signature for the obtained graph by keeping the same signature at each 'old' vertex and making w a zero-vertex. We call this operation an *identification of zero-vertices*. If a pair (G, π) can be obtained from a Shannon triangle G_0 by subdivisions and identifications of zero-vertices, then we say that (G, π) is a *derivative* of the Shannon triangle G_0 (see Fig. 2). It is easily shown that the derivative (G, π) is vertex-parity edge-colorable and $\chi'_p(G, \pi) \leq \chi'_o(G_0)$. Note that the inequality may be strict (as depicted in Fig. 2).

Theorem 3.2. [10, 5] *Every connected proper pair which is not a derivative of a Shannon triangle of type* $(2, 2, 2)$ *or* $(2, 2, 1)$ *is vertex-parity* 4*-edge-colorable.*

Corollary 3.3. [5] *Every proper pair* (G, π) *is vertex-parity* 6*-edge-colorable. Moreover,* $\chi'_p(G, \pi) = 6$ *implies that* (G, π) *is a derivative of a Shannon triangle of type* $(2, 2, 2)$.

In order to characterize the connected proper pairs (G, π) with $\chi'_p(G, \pi) = 6$, we may restrict to pairs with $|\pi^{-1}(1)| = 3$, say $\pi^{-1}(1) = \{a, b, c\}$. For such a pair, we introduce the following specialized notation from [5]: whenever $\{x, y, z\} = \{a, b, c\}$, define

$$G_{xy} = \bigcup_{a,b,c \notin \mathrm{Int}(W_{xy})} W_{xy}\,, \tag{1}$$

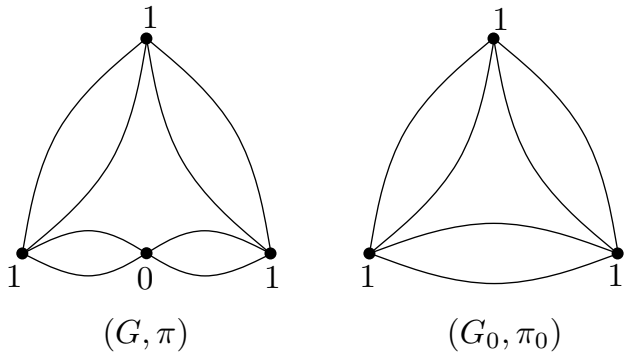


Figure 2: A derivative (G, π) of a Shannon triangle G_0 of type $(2, 2, 2)$. Note that $\chi'_p((G, \pi)) = 4$, while $\chi'_o(G_0) = 6$.

where W_{xy} denotes any xy-walk in G, and in the union each walk is considered as a subgraph of G. Notice that if no such walk W_{xy} exists, then G_{xy} is the null graph. For the intended characterization, we may restrict to the case when none of the subgraphs G_{ab}, G_{bc}, G_{ac} is null. If in addition to this, no two of them share a zero-vertex, then we say that G_{ab}, G_{bc}, G_{ac} are *pairwise internally disjoint*.

Theorem 3.4. [5] *Let (G, π) be a connected proper pair. Then $\chi'_p(G, \pi) = 6$ if and only if the following conditions hold:*

(1) $\pi^{-1}(1) = \{a, b, c\}$.

(2) *G_{ab}, G_{bc}, G_{ac} are pairwise internally disjoint even subgraphs.*

(3) *For every distinct $x, y \in \{a, b, c\}$, x is adjacent to y or at least one $\{x, y\}$-lobe of G_{xy} is not even.*

4 Weak vertex-parity edge-colorings

A similar concept (defined in [6]) to vertex-parity edge-coloring of a pair (G, π) can be introduced as follows. A (not necessarily proper) edge-coloring φ of G is said to be *weak-parity at* a vertex v if at least one color appears at v in parity accordance with π. If φ is weak-parity at each non-isolated vertex of G, then we call it a *weak vertex-parity edge-coloring* of (G, π). Whenever a pair (G, π) admits weak vertex-parity edge-colorings, the minimum number of colors in such an edge-coloring is named the *weak vertex-parity chromatic index* of (G, π), denoted by $\chi'_{\mathrm{wp}}(G, \pi)$.

Proposition 4.1. [6] *The following two conditions are necessary for the existence of $\chi'_{\mathrm{wp}}(G, \pi)$.*

$(w1)$ *For every pendant vertex v of G, it holds that $\pi(v) = 1$.*

$(w2)$ *For no non-empty component H of G, only one $u \in V(H)$ maps to 1 by π, and every other $v \in V(H)$ is of degree 2.*

Whenever the conditions $(w1)$ and $(w2)$ are fulfilled, we say that π is a *weak-proper vertex signature* of G, or equivalently call (G, π) a *weak-proper pair*. Note that each of the components 'forbidden' by $(w2)$ is constructed as follows. For an integer $k \geq 1$,

take vertex-disjoint cycles $C_1, \ldots, C_k$ (1-cycles and 2-cycles are allowed), select one vertex from each cycle and identify the selected vertices into a vertex u. Define a particular vertex signature π_H of the obtained connected graph H by setting $\pi_H(u) = 1$ and $\pi_H(v) = 0$, for every $v \in V(H)\backslash\{u\}$ (see Fig. 3).

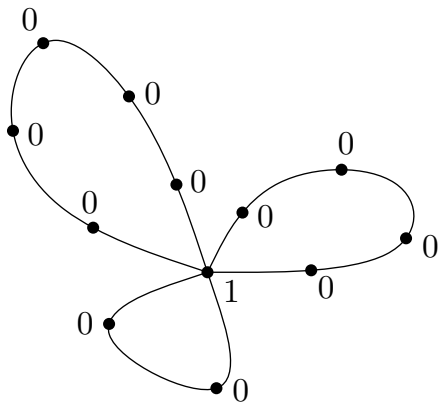


Figure 3: A 'forbidden' component.

Theorem 4.2. [6] *Every weak-proper pair (G, π) is weak vertex-parity 3-edge-colorable.*

4.1 Weak odd edge-colorings of graphs

In the particular instance when the π is the odd vertex signature of a graph G, any weak vertex-parity edge-coloring of (G, π) is referred to as a *weak odd edge-coloring* of G. Then, the parameter $\chi'_{\mathrm{wp}}(G, \pi)$ is written as $\chi'_{\mathrm{wo}}(G)$. According to Proposition 4.1, a necessary and sufficient condition for the existence of a weak odd edge-coloring of G is the absence of vertices incident only to loops. Moreover, by Theorem 4.2, it holds that $\chi'_{\mathrm{wo}}(G) \leq 3$. A complete characterization of G in terms of the value of $\chi'_{\mathrm{wo}}(G)$ is possible, as the following result demonstrates.

Theorem 4.3. [6] *Given a connected graph G whose edge set does not consist only of loops, it holds that*

$$\chi'_{\mathrm{wo}}(G) = \begin{cases} 0 & \text{if } G \text{ is trivial}, \\ 1 & \text{if } G \text{ is odd}, \\ 2 & \text{if } G \text{ has even order or is not even}, \\ 3 & \text{if } G \text{ is non-trivial even of odd order}. \end{cases}$$

4.2 Weak even edge-colorings of graphs

Another particular instance of a weak-proper vertex signature π maps to 1 every pendant vertex of G, and maps to 0 every non-pendant vertex of G. We then call π an *even vertex signature* of G, and any weak vertex-parity edge-coloring of (G, π) is named a *weak even edge-coloring* of G. On this occasion, the parameter $\chi'_{\mathrm{wp}}(G, \pi)$ is written as $\chi'_{\mathrm{we}}(G)$. A complete characterization of G in terms of the value of $\chi'_{\mathrm{we}}(G)$ is also possible.

Theorem 4.4. [6] *Given a connected graph G, it holds that*

$$\chi'_{\mathrm{we}}(G) = \begin{cases} 0 & \text{if } G \text{ is empty}, \\ 1 & \text{if } G \text{ is non-empty with all non-pendant vertices even}, \\ 2 & \text{if } G \text{ has at least one odd non-pendant vertex}. \end{cases}$$

Acknowledgements This work is partially supported by ARRS Program P1-0383 and by Creative Core FISNM-3330-13-500033.

References

[1] P. Erdös, A. W. Goodman, L. Pósa, *The representation of graphs by set intersections*, Canad. J. Math. **18** (1966) 106–112.

[2] F. Jaeger, *Flows and generalized coloring theorems in graphs*, J. Combin. Theory Ser. B **26** (1979) 205–216.

[3] M. Kano, G. Y. Katona, *Odd subgraphs and matchings*, Discrete Math. **250** (2002) 265–272.

[4] B. Lužar, M. Petruševski, R. Škrekovski, *Odd edge coloring of graphs*, to appear in Ars Math. Contemp.

[5] B. Lužar, M. Petruševski, R. Škrekovski, *On vertex-parity edge-colorings*, manuscript (2014).

[6] B. Lužar, M. Petruševski, R. Škrekovski, *Weak-parity edge coloring of graphs*, manuscript (2014).

[7] B. Lužar, R. Škrekovski, *Improved bound on facial parity edge coloring*, Discrete Math. **313** (2013) 2218–2222.

[8] T. Matrai, *Covering the Edges of a Graph by Three Odd subgraphs*, J. Graph Theory **53** (2006) 75–82.

[9] K. R. Matthews, *On the eulericity of a graph*, J. Graph Theory **2** (1978) 143–148.

[10] M. Petruševski, *Odd* 4*-edge-colorability of graphs*, manuscript (2014).

[11] L. Pyber, *Covering the edges of a graph by...*, Graphs and Numbers, Colloquia Mathematica Societatis János Bolyai **60** (1991) 583–610.

[12] P. D. Seymour, *Nowhere-zero* 6*-flows*, J. Comb. Theory Ser. B **30** (1981) 130–135.

[13] P. D. Seymour, *Sums of circuits*, In: Graph Theory and Related Topics (J. A. Bondy and U. S. R. Murty, Eds.), Academic Press, New York (1979) 342–355.

[14] J. Shu, C.-Q. Zhang, T. Zhang, *Flows and parity subgraphs of graphs with large odd-edge-connectivity*, J. Combin. Theory Ser. B **102** (2012) 839–851.

Proceedings of the 6th International Conference on Information Technologies and Information Society ITIS 2014 held in Dolenjske toplice, Slovenia, 5-7 November, 2014
Webpage: http://itis2014.fis.unm.si/
Proceedings edited by: Zoran Levnajić and Biljana Mileva Boshkoska
Proceedings are a collection of contributions presented at the conference

Online free publication
Published by: Faculty of Information Studies, Novo mesto, Slovenia
Main conference sponsor: Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

659.2:004(082)(0.034.2)
316.42:659.2(082)(0.034.2)

INTERNATIONAL Conference on Information Technologies and Information Society (6; 2014; Šmarješke Toplice)

Proceedings [Elektronski vir] / 6th Intenational Conference on Information Technologies and Information Society [also] ITIS 2014, Šmarješke Toplice), 5-7 Novembar 2014; edited by Zoran Levnajić and Biljana Mileva Boshkoska. -El. knjiga. - Novo mesto : Faculty of Information Studies, 2014

ISBN 978-961-93391-3-8 (pdf)
1. Levnajić, Zoran
276897024