Ss. Cyril and Methodius University in Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

2014

Proceedings of the 11th International Conference for Informatics and Information Technology

Held at Hotel Molika, Bitola, Macedonia
11-13th April, 2014

Editors:
Vangel V. Ajanovski
Gjorgji Madjarov

ISBN 978-608-4699-04-0

Conference on Informatics and Information Technology
Web-site: **http://ciit.finki.ukim.mk**
Email: **ciit@finki.ukim.mk**

Publisher:
Faculty of Computer Science and Engineering, Skopje, Macedonia
Ss. Cyril and Methodius University – Skopje, Macedonia
Address: Rugjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, Macedonia

Web-site: http://www.finki.ukim.mk/
Email: contact@finki.ukim.mk

Proceedings Editors:
Vangel V. Ajanovski, PhD, Assistant Professor
Gjorgi Madjarov, PhD, Assistant Professor

Technical editing and cover page: Vangel V. Ajanovski
Conference Photographs: Dejan Mladenovski

Total print run: 150
Printed in Skopje, Macedonia, 2015

ISBN: 978-608-4699-04-0

CIP - каталогизација на публикација
Народна и универзитетска библиотека „Св.Климент Охридски“, Скопје

004.7:621.39(062)
004(062)

PROCEEDINGS of the 11th Conference on Informatics and Information Technology (11 ; 2014 ; Bitola)
Proceedings of the 11th Conference on Informatics and Information Technology : CIIT 2014, April, 11-13, Molika, Bitola, Macedonia / editors Vangel V. Ajanovski and Gjorgji Madjarov. - Skopje : Faculty of Computer Science and Engineering, 2015. - 308 стр. : граф. прикази ; 30 см

Библиографија кон трудовите

ISBN 978-608-4699-01-9
1. Ajanovski, Vangel V. [уредник] 2. Madjarov, Gjorgji [уредник]

Preface

The Conference of Informatics and Information Technology was held for the eleventh time, traditionally in the locality Molika, Bitola, Macedonia in the period April 11-13, 2014. Since 2012 the main organizer of the conference is the Faculty of Computer Science and Engineering (FCSE), located at the Ss. Cyril and Methodius University in Skopje, Macedonia. The FCSE was formed in 2011, as the result of the unification of the two largest institutions in the area of informatics and computer technologies in Macedonia – the Institute of Informatics at Faculty of Natural Sciences and Mathematics and the Institute of Computer Techniques and Informatics at Faculty of Electrical Engineering and Information Technologies. Until 2011, the conference was organized by the Institute of Informatics at the Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius in Skopje. Today, FCSE continues the tradition of giving the opportunity to researchers to present their latest results in the field of Informatics and Information Technologies.

FCSE is the largest faculty in the field of computer science and technologies in Macedonia, and among the largest faculties in that field in the region. At the time of writing FCSE teaching staff consists of 45 professors and 14 teaching assistant and associates. These include many “best in field” persons, such as the most referenced scientist in Macedonia, the most influential professor in ICT industry in Macedonia, etc. The FSCE offers wide range of selective courses to its students making different profiles capable to cope with different professional and scientific challenges ranging from undergraduate (1st cycle) to doctorate programs (3rd cycle). The total number of students enrolled at FCSE is nearly 3000.

New Procedures for Conference Acceptance and Paper Publication

Starting from CIIT 2014, the chairs of the conference together with the help of the Management of the FCSE have introduced a more formal paper acceptance and publication process, in order to improve the quality of the conference, and to broaden the visibility of the conference internationally with the collaboration with the IEEE Computer Chapter Macedonian Section, especially with prof. Marjan Gushev. IEEE support was instrumental for invitation of the keynote lecturers and the organization of the IEEE Best Student Paper Award.

With regards to the acceptance and publication process, CIIT 2014 activities started with the preliminary call for interest in the form of extended abstracts. After the initial call for abstracts, the official Program Committee was formed from invited esteemed researchers from Macedonia. The acceptance for presentation at the conference was sent to the authors of papers that have passed the evaluation by the assigned Program Committee members who decided whether to accept the paper as a full paper, short paper or to reject the papers. In case of conflicting remarks an additional reviews were performed.

Finally, having a presentation live at the conference was a prerequisite for the authors to be invited to submit a final, camera ready version of presented papers. After a long process of revisions of the final versions of the papers, done in collaboration with the authors, 58 of the 67 presented papers have been selected to be included in these Conference Proceedings.

Regarding the Conference Programme

The CIIT 2014 Conference Programme was another step forward. During the official working program of the conference, 2 keynote lectures were given and 67 presentations of accepted full and short papers took place in 10 regular sessions. Also a special student session was coorganized with the IEEE Computer Chapter Macedonian Section for a Best Student Paper Award.

Georgi Dimirovski, Research (retired) Professor of Automation and Systems Engineering at Ss. Cyril and Methodius University of Skopje, Macedonia, Foreign Member of the Academy of Engineering Sciences of Serbia and on a part-time basis guest and invited professor at several universities in Turkey, Hungary, China, Belgium, gave the keynote talk ”Network Q-Learning Controls Prevent Cyber Intrusion Risks: Synergies of Control Theory and Computational Intelligence”.

Dragi Kocev, Post-doctoral researcher, Department of Knowledge Technologies, Jozef Stefan Institute, Slovenia, gave the keynote talk ”Tree ensembles for predicting structured outputs”.

As Conference Chairs of CIIT 2014 and Editors of the Proceedings we hope that the CIIT conference will continue its growth in volume and quality, towards becoming one of the premium venues for presenting current research and development topics in the field of ICT in the region.

Vangel V. Ajanovski and Gjorgji Madjarov.

Figure 1: Official Group Photo in front of Hotel Molika, Bitola, Macedonia

Figure 2: Keynote Lecture by Gjorgji Dimirovski

Figure 3: Keynote Lecture by Dragi Kocev

Figure 4: Weather changes moods in April

Conference Organization

Conference chairs

- Vangel V. Ajanovski, PhD – Assistant Professor, Faculty of Computer Science and Engineering
- Gjorgji Madjarov, PhD – Assistant Professor, Faculty of Computer Science and Engineering

Organizing Committee

- Vangel V. Ajanovski, PhD – Assistant Professor, Faculty of Computer Science and Engineering
- Gjorgji Madjarov, PhD – Assistant Professor, Faculty of Computer Science and Engineering
- Mile Jovanov, PhD – Assistant Professor, Faculty of Computer Science and Engineering
- Magdalena Kostoska, MSc – Teaching Assistant, Faculty of Computer Science and Engineering
- Tomche Delev, MSc – Teaching Assistant, Faculty of Computer Science and Engineering

Local Organization Team

- Emil Stankov, MSc – Associate, Faculty of Computer Science and Engineering
- Bojana Koteska, MSc – Demonstrator, Faculty of Computer Science and Engineering
- Bojan Ilijoski, MSc – Demonstrator, Faculty of Computer Science and Engineering
- Monika Simjanoska, MSc – Demonstrator, Faculty of Computer Science and Engineering
- Bisera Dugalikj, BSc. Eng. – Demonstrator, Faculty of Computer Science and Engineering
- Martin Tashkoski, BSc. Eng. – Demonstrator, Faculty of Computer Science and Engineering
- Mile Kostadinoski, BSc. Eng. – Demonstrator, Faculty of Computer Science and Engineering
- Vladimir Popovski, BSc. Eng. – Demonstrator, Faculty of Computer Science and Engineering

Local Technical Support Team

- Vladislav Bidikov, BSc. Eng. – System Engineer, Faculty of Computer Science and Engineering
- Dejan Mladenovski, BSc. Eng. – System Engineer, Faculty of Computer Science and Engineering

Program committee

- Adrijan Bozhinovski – School of Computer Science and Information Technology, University American College – Skopje, Macedonia
- Aleksandra Mileva – Faculty of computer science, “Goce Delcev” University – Shtip, Macedonia
- Ana Madevska-Bogdanova – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Anastas Mishev – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Andrea Kulakov – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Andreja Naumoski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Boro Jakimovski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Dejan Gjorgjevikj – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Dejan Spasov – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Goce Armenski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Goran Velinov – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Igor Mishkovski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Ivica Dimitrovski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Katerina Zdravkova – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Ljupcho Antovski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Margita Kon-Popovska – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Nevena Ackovska – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Sashko Ristov – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Sasho Gramatikov – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Smile Markovski – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Smilka Janeska Sarkanjac – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Snezhana Cherepnalkovska Dukovska National Bank of the Republic of Macedonia
- Verica Bakeva – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia
- Vesna Dimitrova – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia

Contents

Session 1

Multimedia and Digitization

3D Content Based Medical Image Retrieval: Basic Concepts and Challenges

Katarina Trojacanec, Ivan Kitanovski, Ivica Dimitrovski, Suzana Loshkovska
Department of Software Engineering
Faculty of Computer Science and Engineering, UKIM
Skopje, Macedonia
E-mail: {katarina.trojacanec, ivan.kitanovski, ivica.dimitrovski, suzana.loshkovska}@finki.ukim.mk

***Abstract*— The paper reviews the current status and challenges in 3D Content Based Medical Image Retrieval (CBMIR). Several concepts of the 3D-CBMIR are reviewed, analyzed and discussed. These include volume of interest (VOI) selection/detection/localization, feature extraction from medical volumetric data and similarity measurements. According to the performed research, it should be noticed that there is still room for improvements regarding how they adjust and perform in respect to the clinically relevant volumetric information extraction and utilization.**

Moreover, one of the biggest challenges detected from the performed research is that addressing the condition of the certain patient/case over time, detecting the connection between periodical scans and fusing the information still remains an open question. Retrieving the cases with the same/similar progress of the condition/disorder/disease with the progress for a given patient is a crucial question not yet exhaustively answered in this area.

Keywords—3D-CBMIR, medical volumetric data, VOI, periodical scans, feature extraction, similarity measurements

I. Introduction

A vast amount of medical images are acquired and stored nowadays in medical and research centers. They are one of the most powerful diagnostic techniques. Despite the wide and experienced usage of 2D imaging techniques, many limitations are related to their nature because information extracted from 2D image is not representative, adequate or suffer from information insufficiency in some medical cases.

Taking into consideration the rapid development of the medical imaging techniques, 3D or even more dimensional representations are taking more and more place in the clinical environment. 3D forms are becoming even more valuable factor for the diagnosis and surgical treatment of wide range of pathologies. They enable coherent and collective view, and reduce the guesswork very often experienced with 2D radiology [1, 2].

As a consequence, a vast amount of 3D medical collections have arisen in the medical repositories. Enabling the efficient retrieval and knowledge discovery from such 3D medical collections is becoming dominant in the domain of content based medical image retrieval (CBMIR). Its usage for clinical, research, and/or educational purposes by extracting and manipulating the rich information coming from treating the 3D aspect is a big challenge among researchers in this domain [1, 3, 4, 5, 6].

In general, 3D CBMIR includes several aspects in its basis: volume of interest (VOI) selection/detection/localization, feature extraction from medical volumetric data, and similarity measurements. Research in these aspects has been made towards improvement of the 3D CBMIR, considering the retrieval efficiency, and/or reducing the required time to get the answer from the system. This could induce important benefits in the two possible research directions/goals [4]: clinical decision support [7, 8] and educational purposes [9]. Regarding the application domain, there are studies of the images of healthy people [3] or studies focused on, and specified to different diseases [1, 6]. They could also be categorized on the basis of the body part they treat, such as brain [1, 10], lung [6], heart [8, 11, 12] etc., or the medical imaging technique used for acquisition (e. x. Magnetic Resonance Imaging (MRI) [1, 3, 11, 12, 13], Computed Tomography (CT) [4], High-Resolution Computed Tomography (HRCT) [6], Dual–Energy Computer Tomography (DECT) [7], etc.).

Although the researchers in this domain have published several improvements, the search capabilities of medical CBIR still remain questionable in respect to their effectiveness and efficiency. The reason is mostly related to the specific nature of the medical images and addressing it in the CBIR context. For example, MRI is characterized by limited resolution, intensity inhomogeneity, noise, partial volume effects, leading to geometrical inaccuracies in the model of cortical surface in this case [14]. The specific medical characteristics can affect the basic aspects of the 3D CBMIR.

Choosing appropriate techniques that cover the basic image retrieval aspects and successfully address the research goal is a crucial question. This paper aims to review the current status and widely used methods in the CBIR in the context of 3D medical imaging as well as to detect the room for improvements and to highlight the challenges and future directions.

The paper is organized as follows. Section 2 provides review on the segmentation techniques, their classification, and the validation of the segmentation in the medical context. The current status of the feature extraction algorithms for medical volumetric data is summarized in section 3, while the similarity measurements are discussed in section 4. Section 5 highlights the challenges and future directions in 3D CBMIR.

II. Volume of Interest

The query unit in 3D medical repositories is always a volume of interest (VOI). In some cases the VOI may contain the whole body part (ex. whole brain), while in others, only certain region (ex. certain brain region). It depends on the specific retrieval task, or the question that the retrieval should answer. In the special cases, the VOI might be even a single 2D image slice [3].

The detection of volumes of interest or at least the identification of organs in the body can be an important first step to focus better and analyze the data of interest and extract automatically semantic labels from the visual image information [15]. Focusing the retrieval process on a VOI from the image is considered as a very important aspect of a practical retrieval system [13]. Moreover, it is of great importance for localized monitoring of the pathology or changing state of certain specific body part/region that usually reflects the progression of the disease/abnormality, from one side, or the normal state change, from the other side. Thus, the delineation of anatomical structures, and other regions/volumes of interest, i.e. which by definition is referred to as image segmentation is a big challenge and a key component of image analysis and interpretation [16].

A. *Classification*

Several classifications of the image segmentation techniques are common in the literature. Some of them [17, 18, 19] are structured and depicted on figure 1. Most of the methods belong to more than one class viewed and some class might be considered as a subclass of other classes according to the type of classification.

Segmentation approaches in general, can be classified as manual outlining methods, semi-automated methods, and fully automated methods. The manual segmentation is characterized by several drawbacks such as the large amount of work required, and the considerable inter- and intra-rater variability of the results. Thus semi- and fully automatic segmentation methods are preferred [17].

From the point of view of the principal techniques, the segmentation algorithms can be classified into three categories: algorithms based on threshold, clustering techniques, and deformable models [18].

According to the approach used for segmentation, another, wider and more comprehensive classification of the segmentation techniques for medical volumes is proposed in [19]. According to this, three classes are distinguished: structural techniques, stochastic techniques, and hybrid techniques.

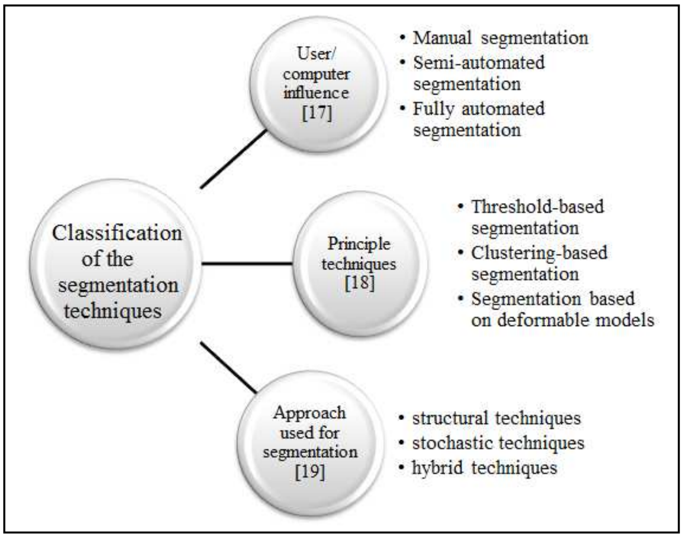

Classification of the segmentation techniques
User/ computer influence [17]
• Manual segmentation
• Semi-automated segmentation
• Fully automated segmentation
Principle techniques [18]
• Threshold-based segmentation
• Clustering-based segmentation
• Segmentation based on deformable models
Approach used for segmentation [19]
• structural techniques
• stochastic techniques
• hybrid techniques

Fig. 1. Different classifications of the segmentation techniques

The application domain of the segmentation techniques that relies on medical imaging is very critical, due to the complexity and variability of the region/volume of interest. Some of the reasons are normal anatomic variation, post-surgical anatomic variation, vague and incomplete boundaries, inadequate contrast, artifacts and noise [20]. For example, critical parts that the segmentation of brain gliomas has to address include the infiltration of cells into the tissue, inducing unsharp borders with irregularities and discontinuities (a tumor is not necessary a single connected object), the great variability in their contrast uptake (depending on their vascularisation), and their appearance on standard MRI protocols [21]. Moreover, it is important to stress that the same organs or structures may appear different in different slices or imaging modalities, leading to the necessity of appropriate and distinctive segmentation method [18].

Taking into account that there is no unique segmentation technique that can reach all different specifics imposed by different domains and produce the satisfactory results for all of them [22], the segmentation methods are optimized into different directions considering the medical application domain:

- Specific imaging modalities, such as MRI, CT, etc.,
- Specific anatomic structures, such as the brain, the lungs etc.,
- Specific disease/abnormality, such as brain tumor, brain atrophy, etc.

B. *Validation*

The validation of medical image segmentation is crucial and one of the most critical parts in the medical applications of the image processing techniques. A very important concern is the selection of reference segmentations. Moreover, the validation of the segmentation in the medical context needs to be conducted together with the clinical relevancy and impact [23], which is very domain and objective specific. The authors of [23] provide a recent review of the evaluation methods for medical image segmentation, stressing that in general, a set of performance measurements should be provided for a complete evaluation and comparison of segmentation techniques. Additionally, they propose a novel concept of accuracy assessment in medical image processing which is easily extensible to any kind of processing and imaging modality, making it very versatile and flexible.

One important criterion here is also a computation time [23, 24]. According to [24], the current standard is about a few minutes regarding the processing time.

C. Related Work and State-of-the-art

A good overview on the 3D segmentation is provided in [20], as an example of the application of machine learning techniques in radiology. Another review on medical image segmentation methods is provided by [18]. They also provide examples of the application of the reviewed techniques to female pelvic cavity. Discussion about medical image segmentation is offered in [25] by identifying three generations of the segmentation algorithms. The authors provide a representative set of examples, available software, and reference databases. A survey on several fundamental segmentation techniques with the application to medical imaging is given by the authors of [26]. They focus on the segmentation methods for general purpose and not specific to certain anatomic organ, that are easily extensible to 3D which can flexibly make use of statistical information, and belong to one of the four categories: region-based, boundary-based, hybrid, and atlas-based. Another review on segmentation techniques used for medical volumes is given in [19] where the techniques are categorized in three main groups: structural, statistical, and hybrid techniques.

Several surveys/reviews are published by focusing on a specific class of segmentation techniques and/or a specific application domain. For instance, a survey on atlas-based image segmentation is provided in [27]. It covers the studies that discuss some the important steps and questions from the field of atlas based segmentation. Their domain of interest includes medical images, with special emphasis on cardiac images. In [28], a review on statistical shape models for 3D medical image segmentation is given and a survey of applications in the medical field and a discussion of future developments are provided. The automatic segmentation methods used for multiple sclerosis lesion segmentation are systematically reviewed in [29].

On the other hand, there are a significant number of studies that conduct a review or propose new segmentation methods on the bases of the application domain. For instance, a comprehensive review on automated techniques for multiple sclerosis lesion detection is provided by [30]. The fully automated method for diffuse white matter lesion segmentation in brain MRI is proposed in [17] in which lesion voxels are detected as outliers of the intensity distribution of normal tissues in multi-sequence MR images. Rating the significance of lesion detection using the random field theory is a very important contribution of the research conducted in [17]. With the aim to take into consideration the large variability and the contrast differences of the prostates, the authors of [31] propose a new automatic atlas-based segmentation method for the segmentation of the prostate MR images. They have obtained satisfactory results, but have suggested some critical points leaving the room for improvements. Another robust and fully automatic approach for segmentation of prostate MRI is proposed in [32]. The method is based on using a probabilistic atlas and a spatially constrained deformable model.

However, there are some structures that have some features in common. Such an example includes lung nodules, liver metastases, and lymph nodes. They are mostly homogeneous and roughly spherical or ellipsoid-shaped objects. They are also often located close to other structures that exhibit a similar density in CT images [33]. The characteristics of this structures lead to the expectation for good results if the same, proper segmentation technique is used in all three cases. The authors in [33] used a hybrid algorithm originally developed for lung nodules and combining the threshold based approach model-based morphological processing.

It should be noticed that although in some cases for segmentation of medical volumetric data, the fundamental techniques might be useful [26], there are plenty of cases where specific domain knowledge has to be incorporated in a certain way, leading to a strictly problem oriented segmentation methods. Later techniques are mostly promising ones in the medical domain, due to the sensitivity of the domain and some challenging characteristics of the medical volumetric data imposed by the modality, acquisition process, or the VOI that is subject of the segmentation.

III. Feature Extraction

Extraction of robust and precise visual features from medical images is a critical task. It is even more challenging to enable reach and complete information extraction and representation from 3D medical context. Although the research has shown that the more visual features are used the better the results are 3D or more dimensional images impose the necessity for optimization and compactness in this context [13, 15].

There are image features that are likely to be the most relevant from the perspective of human image interpretation. Some of the key features include lesion shape, boundaries, density or intensity, presence or absence of enhancement with intravenous contrast material, texture, and whether a lesion is solitary or multiple. It should be noticed that in general, the relative importance of features will vary across modality and disease targets. [34]. Thus, the selection of the feature extraction algorithms should be target oriented rather than generic (expected to represent good enough the visual information from the images of different modalities and with different kinds of targets within the image itself).

Wide range of studies includes evaluation of the feature extraction methods. Several descriptors (or their variations) are commonly used in these studies. The discussion is given in the next paragraphs.

Different variations of the co-occurrence matrices adjusted to the application domain have been evaluated in [1, 3, 35].

The authors from [3] are working on examination of the descriptor in sense of capturing the differences associated with gender (which in most cases are almost inconsiderable). They are using multi-sort, multi-dimensional co-occurrence matrices as very sensitive descriptors to tenuous differences in brain image patterns and reflection/rotation invariant, which is very important characteristic for their application domain because of the specific "reflected" intensity distribution in the left and right hemispheres and unpredictable sulcal variability [3]. A suited version of grey level co-occurrence matrices to 3D context is used in [1]. It is then applied to brain MRI with and without volume detection. An extended co-occurrence matrices for 3D texture analysis is used in [35] too, addressing three important aspects for their research: consideration of the change of original image data dimensionality from 2D to 3D, increasing the sensitivity and specificity of co-occurrence descriptors, and rotation and reflection invariance. It should be noticed that the problem of choosing basic features (matrix axes) depends on the application domain, namely the image data modality and the specific analysis to be performed [35].

It can be summarized that this descriptor is suitable to detect small, even unnoticeable differences and invariances in respect to the rotation and/or reflection. However, when compared to other texture descriptors extended to fit the 3D representations, such as 3D Local Binary Patterns (3D LBP), 3D Wavelet Transforms (3D WT), and 3D Gabor Transforms (3D GT), it is outperformed by the others [1]. In fact, the authors of [1] present the results from the retrieval process using their texture-based online system for 3D images, MIRAGE on the bases of the mean average precision (MAP), and query time with and without VOI detection. The application domain includes 3D brain MRI. According to the presented results obtained by authors, 3D LBP (Local Binary Patterns) descriptor outperforms the other considering the precision and the query time. Additionally, it can be noticed that if the VOI selection is included, the query time is 4 times faster. The good discriminative power of LBP descriptor explains the obtained results. Moreover, another advantage of this descriptor is the simplicity of its implementation.

The benefits from using localized low-level features for CBIR of HRCT images of the lung are evaluated in [6]. In fact, the authors exploit the location of the abnormal/pathological tissue in the lung in the system, thus leading to the improvement in terms of early retrieval precision in comparison with the approach based on global features only.

A 3D SIFT descriptor is introduced in [36] with the characteristic to encode the information local in both space and time in a manner which allows robustness to orientations and noise. It is introduced for video or 3D imagery such as MRI data, and evaluated on action classification in a video. In the performed comparative analysis with other descriptors, the proposed 3D SIFT outperforms the others.

IV. Similarity Measurements

To retrieve the images in a CBIR, the user submits the query image example to the system. The same feature extraction algorithm (as for the 3D images in the database) is applied to the query. Then the similarity between the visual image representation of the query image and the representations of all images in the database is determined. The sorted list of the database images ordered by their similarity with the most similar at the top is returned as a result.

Based on the empirical estimates of the feature distribution, different similarity measurements have been used to make a similarity comparison, namely, Minkowski-Form distance, which is a generic form of the widely used Euclidean distance, Mahalanobis distance, quadratic form distance, proportional transportation distance, earth mover's distance, Kullback-Leibler divergence and Jeffrey divergence, etc. [37].

For example, histogram intersection is used as a similarity measurement in [38] in the case of the feature extraction performed by the LBP descriptor. Euclidean distance is also very common similarity measurement. It is used in the case of 3D Grey Level Co-occurrence Matrices, 3D Wavelet Transform, and 3D Gabor Transform descriptors in [38].

On the other hand, if the graph-based representation of the features is used rather than vector-based, special computational methods for similarity assessment, referred to as graph matching, are required [34].

There are cases, especially for diagnostic purposes, where subtle geometrical differences between imaged structures may be very important. In these cases, the similarities are defined through the notion of elastic deformations required to transform one shape into another are useful. The energy needed to transform one shape into another is assumed to be inversely proportional to similarity [34].

Another, very promising approach for measuring the similarity is by using statistical classifiers that classify new instances using high-level information extracted from the training set of instances. It is clear that the labels have to be known for the training set instances. This approach is thus very problem oriented [34].

Choosing a metric to measure the similarity has direct influence on the performance of the CBIR [37]. It is dependent on the type of the descriptors as well as on their representation [34].

V. Challenges in CBIR for 3D Medical Imaging

Although many studies have been performed in the field of image retrieval, various aspects related to the application of CBIR in 3D medical imaging are still open for research and improvement. They are highlighted in the following paragraphs.

Several challenges are identified in the context of medical image segmentation:

- Robustness on the difficulties imposed by image acquisition process such as noise, intensity inhomogeneities, partial volumes, ill-defined boundary, low contrast etc.
- Capability to properly delineate the relevant structure (tissue, pathology, anatomical structure…)

- Decreased user/expert influence
- Degree of the domain knowledge usage
- Complexity and processing time

The segmentation techniques studied and to be proposed in the future should tend to address as much as possible of these challenges to reach the desired usefulness and precision in this application domain.

The focus in feature extraction from 3D medical images should be directed to precisely extract as much as possible information directly from the volume context. Moreover, taking into consideration the spatial relationships is another, very big challenge. Moreover, it is considered as very close to the physician's mental model of the patient, and very important for the diagnosis and prognosis.

The appropriateness of the similarity/dissimilarity measurement and the comparison performed between the query unit and all units in the database is also very important and should be carefully chosen.

Moreover, it is very important to notice that most of the studies suffer from leak of treating the periodical connection between medical examinations. This is very important in the context of addressing the condition of the patients regarding the serial scans or the progression of some diseases.

Finding the cases with the same/similar progress of the condition/disorder/disease with the progress for a given patient remains an open important question that CBMIR has to be able to answer. For example, a few queries we found challenging in this context are:

- Find all cases with disease progression similar to the query one
- Find all cases with certain condition that did not progressed to a certain disease for the examined period similar to the query one
- Find all cases with the similar treatment reaction over the period similar to the query one

Answering these questions might be helpful from different aspects, including diagnosis and prognosis support, educational purposes, completion of the general image about the patient state in the case of missing one or more scans of his/her study, proposing the treatment and/or monitoring the patient's reaction on particular treatment.

VI. Conclusion

Review on the current status of the 3D CBMIR was provided in the paper regarding its crucial concepts such as segmentation of the VOI, feature extraction from medical volumes, as well as the similarity/dissimilarity measurements. As a result, the advantages and disadvantages of the current and mostly used techniques were stressed and their suitability for application to 3D medical imaging domain was discussed.

According to the research, it is noticed that there is still room for improvement. The aspects related to the application of CBIR in 3D medical imaging that still remains open for research are detected. Thus, several challenges were highlighted and possible future directions proposed.

Regarding the 3D segmentation, the focus should be on addressing the robustness to the image acquisition specificities, capability to properly delineate the relevant structure, decreasing the user/expert involvement, the degree of the domain specific knowledge usage, and the complexity and the required processing time. From the point of view on the feature extraction process, the comprehensive information precisely extracted directly from the medical volumes as well as treating the spatial relationships should be the crucial subject of consideration in the future researches. Considering the similarity/dissimilarity measurements, the proper way for comparison should be taken. Moreover, one of the biggest challenges that have arisen is involving and addressing the time progression and periodical connections into the retrieval process.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University" within the project "Scalable Image Annotation".

References

[1] G. Xiaohong, Y. Qian, M. Loomes, R. Comley, B. Barn, A. Chapman, J. Rix, R. Hui, and Z. Tian, "Retrieval of 3D medical images via their texture features." International Journal On Advances in Software vol. 4, no. 3 and 4, 2012, pp:499-509.

[2] [http://www.massgeneral.org/imaging3d/benefits/] last visited: 22.02.2014

[3] V. Kovalev, F. Kruggel, "Retrieving 3D MRI Brain Images", In: A.W.M. Smeulders (ed.), VISIM Workshop, Information Retrieval and Exploration from Large Medical Image Collections, pp. 53-56. Utrecht, The Netherlands 2001.

[4] A. Foncubierta-Rodríguez, H. Müller, A. Depeursinge, "Region-based volumetric medical image retrieval",. Proc. SPIE 8674, Medical Imaging 2013: Advanced PACS-based Imaging Informatics and Therapeutic Applications, 867406, 2013.

[5] Y. Qian, X. Gao, M. Loomes, R. Comley, B. Barn, "Content-based Retrieval of 3D Medical Images", The Third International Conference on eHealth, Telemedicine, and Social Medicine (eTELEMED 2011), pp. 7-12, 2011.

[6] A. Depeursinge, T. Zrimec, S. Busayarat, and H. Müller, "3D lung image retrieval using localized features", In SPIE Medical Imaging, International Society for Optics and Photonics, 2011, pp. 79632E-79632E.

[7] A. Foncubierta–Rodríguez, A. Vargas, A. Platon, P. A. Poletti, H. Müller, and A. Depeursinge. "Retrieval of 4D dual energy CT for pulmonary embolism diagnosis." In Medical Content-Based Retrieval for Clinical Decision Support, pp. 45-55. Springer Berlin Heidelberg, 2013.

[8] L. C. C. Bergamasco, and F. L. Nunes, "Content Based Retrieval for 3D Medical Models: A Study Case Using Magnetic Resonance Imaging" Data Inicio, 2012.

[9] A. Rosset, H. Muller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib, "Casimage project - a digital teaching files authoring environment", Journal of Thoracic Imaging 19(2), 2004, pp. 1-6

[10] S. Yang, L. Shapiro, M. Cunningham, M. Speltz, C. Birgfeld, I. Atmosukarto, and S. I. Lee, "Skull retrieval for craniosynostosis using sparse logistic regression models", In Medical Content-Based Retrieval

for Clinical Decision Support, pp. 33-44, Springer Berlin Heidelberg, 2013.

[11] T. Glatard, J. Montagnat, and I. E. Magnin, "Texture based medical image indexing and retrieval: application to cardiac imaging." In Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, pp. 135-142. ACM, 2004.

[12] L. C. Bergamasco, F. L. Nunes, "Applying Distance Histogram to retrieve 3D cardiac medical models". In AMIA Annual Symposium Proceedings, Vol. 2013, p. 112, American Medical Informatics Association, 2013.

[13] K. Simonyan, M. Modat, S. Ourselin, D. Cash, A. Criminisi, A. Zisserman, "Immediate ROI search for 3-d medical images". In Medical Content-Based Retrieval for Clinical Decision Support, pp. 56-67, Springer Berlin Heidelberg, 2013.

[14] S. Osechinskiy, F. Kruggel, "PDE-based reconstruction of the cerebral cortex from MR images". IEEE International Conference on Engineering in Medicine and Biology (EMBC'10), 2010, pp. 4278-4283.

[15] H. Müller, and H. Greenspan, "Overview of the Third Workshop on Medical Content–Based Retrieval for Clinical Decision Support" (MCBR–CDS 2012). In Medical Content-Based Retrieval for Clinical Decision Support, pp. 1-9. Springer Berlin Heidelberg, 2013.

[16] K. Van Leemput, D. Vandermeulen, F. Maes, S. Srivastava, E. D'Agostino, and P. Suetens. "Model-Based Brain Tissue Classification." In Handbook of Biomedical Image Analysis, pp. 1-55. Springer US, 2005.

[17] F. Yang, Z. Shan, F. Kruggel, "White matter lesion segmentation based on feature joint occurrence probability and χ2 random field theory from magnetic resonance (MR) images", *Pattern Recognition Letters*, *31* 9, 2010, pp. 781-790.

[18] Z. Ma, J. M. R. Tavares, R. N. Jorge, T. Mascarenhas, "A review of algorithms for medical image segmentation and their applications to the female pelvic cavity." Computer Methods in Biomechanics and Biomedical Engineering 13, no. 2, 2010, pp. 235-246.

[19] S. Lakare, and A. Kaufman. "3D segmentation techniques for medical volumes." *Center for Visual Computing, Department of Computer Science, State University of New York*, 2000.

[20] S. Wang, and R. M. Summers. "Machine learning and radiology." *Medical image analysis* 16, no. 5, 2012, pp. 933-951.

[21] E. D. Angelini, O. Clatz, E. Mandonnet, E. Konukoglu, L. Capelle, and H. Duffau. "Glioma dynamics and computational models: a review of segmentation, registration, and in silico growth algorithms and their clinical applications." Current Medical Imaging Reviews 3, no. 4, 2007, pp. 262-276.

[22] A. A. Farag, N. Mohamed, A. A. El-Baz, and H. Hassan. "Advanced segmentation techniques." In *Handbook of biomedical image analysis*, pp. 479-533. Springer US, 2005.

[23] F. F. Pizzorni, and G. Menegaz. "Performance Evaluation in Medical Image Segmentation." *Current Medical Imaging Reviews* 9, no. 1, 2013, pp. 7-17.

[24] S. Bauer, R. Wiest, L. P. Nolte, and M. Reyes. "A survey of MRI-based medical image analysis for brain tumor studies." *Physics in medicine and biology* 58, no. 13, 2013, R97.

[25] D. J. Withey, and Z. J. Koles. "Medical image segmentation: Methods and software." In *Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging, 2007. NFSI-ICFBI 2007. Joint Meeting of the 6th International Symposium on*, pp. 140-143. IEEE, 2007.

[26] Y. C. Hu, M. D. Grossberg, and G. S. Mageras. "Survey of recent volumetric medical image segmentation techniques." *Biomed Eng* 2009, 321-346.

[27] H. Kalinic, "Atlas-based image segmentation: A Survey." *Croatian Scientific Bibliography*, 2009.

[28] T. Heimann, and H. P. Meinzer. "Statistical shape models for 3D medical image segmentation: A review." *Medical image analysis* 13, no. 4, 2009, pp. 543-563.

[29] D. G. Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins. "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging."*Medical image analysis* 17, no. 1, 2013, pp. 1-18.

[30] X. Lladó, O. Ganiler, A. Oliver, R. Martí, J. Freixenet, L. Valls, J. C. Vilanova, L. R. Torrentà, and À. Rovira. "Automated detection of multiple sclerosis lesions in serial brain MRI." *Neuroradiology* 54, no. 8, 2012, pp. 787-807.

[31] A. G. Merida, R. Marti. "Atlas based segmentation of the prostate in MR images." In *MICCAI: Segmentation Challenge Workshop*. 2009.

[32] M. Sébastien, J. Troccaz, and V. Daanen. "Automated segmentation of the prostate in 3D MR images using a probabilistic atlas and a spatially constrained deformable model." *Medical physics* 37, no. 4, 2010, pp. 1579-1590.

[33] J. H. Moltz, L. Bornemann, J. M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte et al. "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans." *Selected Topics in Signal Processing, IEEE Journal of* 3, no. 1, 2009, pp. 122-134.

[34] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content-based image retrieval in radiology: current status and future directions". Journal of Digital Imaging, 24(2), pp. 208-222, 2011.

[35] V. A. Kovalev, F. Kruggel, H. J. Gertz, D. Y. von Cramon, "Three-dimensional texture analysis of MRI brain datasets". Medical Imaging, IEEE Transactions on, 20(5), 2001, pp. 424-433.

[36] P. Scovanner, S. Ali, M. Shah. "A 3-dimensional sift descriptor and its application to action recognition". In Proceedings of the 15th international conference on Multimedia, 2007, pp. 357-360, ACM.

[37] T. W. Cai, K. Jinman, and D. D. Feng. "4 CONTENT-BASED MEDICAL IMAGE RETRIEVAL." 2008, pp. 83-113.

[38] X . Gao, Y. Qian, R. Hui, M. Loomes, R. Comley, B. Barn, A. Chapman, and J. Rix. "Texture-based 3D image retrieval for medical applications." *IADIS e-Health2010*, 2010, pp. 29-31.

Graphical user interface for image restoration

Ana Ljubotenska, Igor Stojanovik
Department of Computer Engineering and Communication Technologies, Faculty of Computer Science
'GoceDelcev' University
Stip, Macedonia
e-mail: {ana.ljubotenska, igor.stojanovik}@ugd.edu.mk

***Abstract* — The images are created to display or record useful information. Due to the large number of errors that exist in the procedure of image processing, the recorded image is always a degraded version of the original image. There are various ways in which the image can be degraded, but this study only examined blur caused by the movement of camera or scene being recorded. The restoration of images refers to the reconstruction or estimation of images with noise and blur, i.e. the operation on the image which is inverse on the imperfections which occurs during the system formatting of the image. This paper presents a graphical user interface (GUI) for image restoration, developed in MATLAB. The original image is degraded by setting the parameters of blur. The restoration of degraded image is performed by using linear Wiener filter and nonlinear Lucy-Richardson filter for image restoration. The aim is to examine the effectiveness of these filters, with assessment and analysis of the quality of the restored images by examining several important measures for image quality, compared to their ideal values.**

Keywords— Graphical User Interface; image restoration; noise; blur; Wiener filter; Lucy-Richardson filter; image quality measurements

I. Introduction

Graphical user interface (GUI) is defined as a graphical presentation of one or more windows that contain controls, called components, which allow the user to perform interactive tasks. Using MATLAB, solutions to many problems are formulated, involving the matrix representation in a part of time. Such are the problems connected with image processing. Developing GUI using MATLAB can take place in one of the two following ways:

- By using the Graphical User Interface Development Environment, contracted GUIDE, which is a tool for the construction of interactive GUI;
- By creating files with code which generate GUI as scripts or functions. This mode is also known as a program construction of GUI.

For the purposes of this paper, the first approach is used. It begins by creating a figure that is filled with components that are selected from the appropriate graphic layout editor. GUIDE adds a file with code which refers to the feedback functions for the graphical interface and its components. GUIDE keeps the figure or the design as a FIG-file and a file with code, which is known as M-file. For each created FIG file when it is saved M file is also created with the same name as the FIG-file. In performing the GUI, with the opening any of these two files the other one opens at the same time. Each M file consists of two basic functions for each created GUI component. These functions are CreateFcn and Callback. MATLAB as a High-performance language, by combining the calculations, visualization and programming is an environment that is easy to use for the image processing. Image Processing Toolbox, contracted IPT, is a collection of standard functions, algorithms and applications for image processing, analysis and virtualization [1]. They provide image enhancement, its blurring, noise detection and reduction, image segmentation, geometric transformations or image restoration. The three major topics in the area of image processing are: image restoration [2], improving image quality and compression of images. In this GUI, only the restoration of images is concerned [3]. The restoration of images is current in the field of research and discovering applications in many fields, such as medical visualization, spatial snapshots, forensic science and commercial images.

II. Modeling of the process of the image degradation

The basic scheme of image restoration is given below.

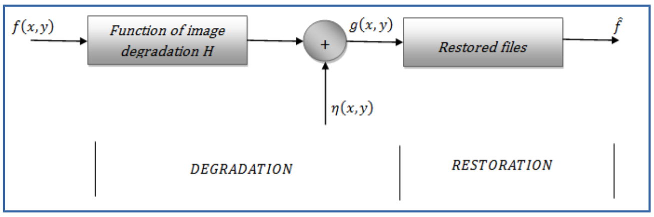


Fig. 1 Model of degradation / restoration of the image

Here $f(x,y)$ is the original image, $g(x,y)$ is the degraded image, $\hat{f}$ is the restored image, and $\eta(x,y)$ is noise. The figure shows that the process of restoration precedes the degradation process that occurs by applying a certain function to degrade the original image, to which a noise can be added [4]. In tests the noise is not taken into account. In this way a degraded image is obtained, on which certain filters for image restoration can be applied. As a result a restored image is obtained. Restoration attempts to perform an operation on an image that got imperfections during the formatting process of the image. The more the restored image is similar to the original image, the process of restoration is more successful. This model is applied to the created GUI, where firstly the image is loaded, then it is degraded and restored using Wiener and Lucy-Richardson filters.

The model of degradation of images mathematically is representing the dependence on the input image $f(x,y)$ to get a degraded image $g(x,y)$. This dependence is mathematically defined by:

$$g(x,y)=h(x,y)*f(x,y)+\eta(x,y) \quad (1)$$

Here h is a factor of disturbance, which is still known as the point spread function or abbreviated PSF [5]. In the spatial domain, PSF describes the extent to which in the optical system of blur a point of light is spreading. PSF is actually the inverse Fourier transformation of the optical transfer function (OTF).

This paper examines the degradation which is carried out only by blurring the image. Blurring is the form of a reduction of the bandwidth of an ideal image, due to imperfections in the image formation. It can be caused by: relatively motor movement between the camera and the original scene, optical system that is out of focus or atmospheric turbulence.

There is a test with a linear motion blurring which occurs when the scenes are recorded with a camera that has a fixed position, with constant speed $v_{relative}$, at an angle θ radians, towards the horizontal axis in the interval [0, $t_{exposure}$]. Then the disorder is one – dimensional. If the length of the movement is labeled with L and calculated as $L=v_{relative}t_{exposure}$, than we have PSF as:

$$H(x,y;L,\theta)=\begin{cases} \frac{1}{L}, if\sqrt{x^2+y^2}\leq\frac{\frac{L}{2}\wedge x}{y}=-tan\theta \\ 0, other \end{cases} \quad (2)$$

In the created GUI, the blurring is done by setting the parameters: length and angle of the blur. They are set in section blurs parameters, by setting the value directly or via a slider. The length of blurring is between 1 and 150, and the angle is in the range from 0 to 180 degrees. Blurring the images is modeled by the use of IPT function fspecial, as follows:

PSF = fspecial (‘motion’, len, theta) (3)

This function call, returns PSF which approximates the effects of linear motion of the camera on pixels with length len.

III. Methods for image debluring

Deblurring images is an interactive process. In order to fully clean blurred image, sometimes is necessary to apply the demisting process several times, depending on the specified parameters in each iteration. In this paper demisting image is done by using two filters for image restoration: Wiener and Lucy-Richardson. Wiener [6] filter is the best known approach to linear image restoration. He is also direct one, which means that the solution is obtained through one application of the filter. This method attempts to estimate $\hat{f}$, which minimize the function of static error:

$$e^2=E\left\{\left(f-\hat{f}\right)^2\right\} \quad (4)$$

where E is the expected value operator, $\hat{f}$ is the estimated value of f, i.e. indicates the behavior of the filter in the coordinates (x,y), or it is restored image, and f is the original, non-degraded image. Static error is still known as the mean-squared error (MSE). The solution of this equation in the frequency domain is:

$$\hat{F}(x,y)=\left[\frac{1}{H(x,y)}\frac{\left|H(x,y)\right|^2}{\left|H(x,y)\right|^2+\frac{S_\eta(x,y)}{S_f(x,y)}}\right]G(x,y) \quad (5)$$

where $H(x,y)$ is a function of degradation, $G(x,y)$ is Fourier transformation of degraded image, $\left|H(x,y)\right|^2=H^{\square}(x,y)H(x,y)$, $H^{\square}(x,y)$ is $H(x,y)$ complex conjugate, $S_\eta(x,y)=\left|N(x,y)\right|^2$ is the spectral power of noise and $S_f(x,y)=\left|F(x,y)\right|^2$ is spectral power of original, non-degraded image. The ratio $\frac{S_\eta(x,y)}{S_f(x,y)}$ is known as the rate of intensity noise - signal.

The implementation of the Wiener filter in MATLAB with IPT using a function deconvwnr with the form:

fr = deconvwnr (g, PSF) (6)

Lucy- Richardson filter is nonlinear [7], iterative technique [8] for image restoration. The mathematical form is obtained from the following equation, which should satisfy the requirement for convergence in each iteration:

$$\hat{f}_{k+1}(x,y)=\hat{f}_k(x,y)\left[h\frac{(-x,-y)*g(x,y)}{h(x,y)\hat{f}_k(x,y)}\right] \quad (7)$$

where $\hat{f}$ denotes the value of restored image, g denotes the degraded image, and h is a function of degradation. As with all other nonlinear methods, generally is difficult to answer the question of when to stop the L-R algorithm. The method can consider the obtained outputs and stop when the result in a given application will be accepted or perform for the exact number of iterations, as in GUI in this paper, where up to 1000 iterations are possible, but initially 20 iterations should be taken. This filter can be implemented in the IPT with the function deconvlucy whose basic syntax is:

fr = deconvlucy (g, PSF, NUMIT, DAMPAR, WEIGHT)(8)

where fr is restored image, g is degraded image, PSF is a function of the point spread, NUMIT is the number of iterations, DAMPAR is a scalar that specifies the threshold deviation of the resultant image of degraded image and WEIGHT is the array with the same size as g and measure the width of each pixel to reflect its quality.

IV. Image quality measurements

To determine the effectiveness of the process of image restoration, there are calculations of several measures for image quality, such as: average absolute difference – AAD, maximum difference – MD, mean square error – MSE, normalized absolute error – NAE, structural content – SC, normalized cross correlation – NCC, signal to noise ratio – SNR, peak signal to noise ratio – PSNR and others [9].

MSE between restored image $\hat{f}(x,y)$ and the original image $f(x,y)$, with the size N x M, for each pixel is the average error:

$$MSE=E\left\{(f-\hat{f})^2\right\}=E\left[\left|f(x,y)-\hat{f}(x,y)\right|^2\right] \quad (9)$$

The ideal value is 0. If it is a spatial domain, where $f(x,y)$ is a random process, then mean – square error has the form:

$$MSE=\frac{1}{MN}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left|f(x,y)-\hat{f}(x,y)\right|^2 \quad (10)$$

Signal to noise ratio (SNR) for restored image is calculated as:

$$SNR_{\hat{f}}=10\log_{10}\left(\frac{f(x,y)}{\hat{f}(x,y)-f(x,y)}\right) \text{ (dB)} \quad (11)$$

Appropriate for degraded image is calculated by:

$$SNR_g=10\log_{10}\left(\frac{f(x,y)}{g(x,y)-f(x,y)}\right) \text{ (dB)} \quad (12)$$

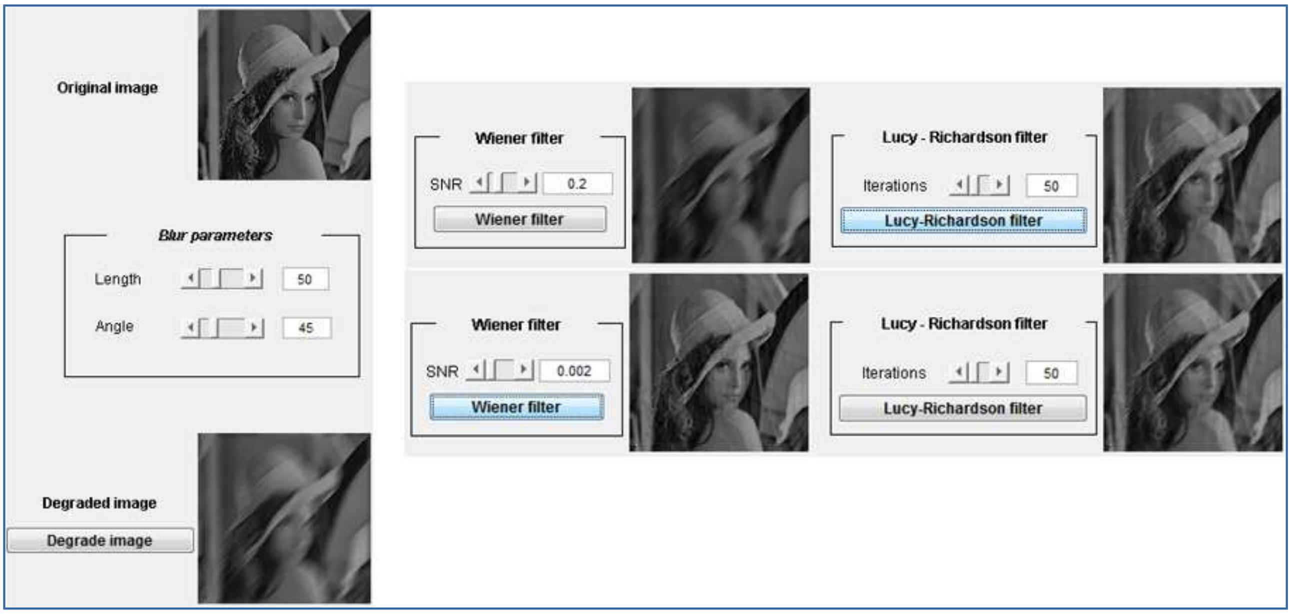
Ideal values are the higher ones. Other image quality measurements that are part of the GUI are calculated according to equations given in table 1.

Table 1 Image quality measurements

Image quality measurements	Abbreviation	Formula
Normalized Cross - Correlation	NCC	$NCC=\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y)\hat{f}(x,y)/$
Average Difference	AD	$f(x,y)-\hat{f}(x,y)$ $(\|)/NM$ $\sum_{y=1}^{M}\square$ $AD=\sum_{x=1}^{N}\square$
Structural Content	SC	$SC=\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y)^2/\sum_{x=1}^{N}\sum_{y=1}^{M}\hat{f}$
Maximum Difference	MD	$MD=max(\|f(x,y)-\hat{f}(x,y)\|$
Normalized Absolute Error	NAE	$NAE=\sum_{x=1}^{N}\sum_{y=1}^{M}\|f(x,y)-\hat{f}(x,$

V. Experimental results

Created GUI analysis was done for several values of the blur parameters, SNR and L-R iterations. In the first experiment blur parameters are: Length 50 and Angle 45. Blurred image with this values is given on figure bellow. Image first is restoring with Wiener filter with SNR 0,2, and then with L-R filter in 50 iterations. The restored image with Wiener filter is far from the original, i.e. it's more blurred though restored image with L-R, which is more comparable to the original image i.e.it is clearer. If value for SNR is 0,002, and other parameters leave the same, the restored image with Wiener filter is comparable to the original. This can be visually observed from the results shown in figure 2.



Length 50

Angle 45

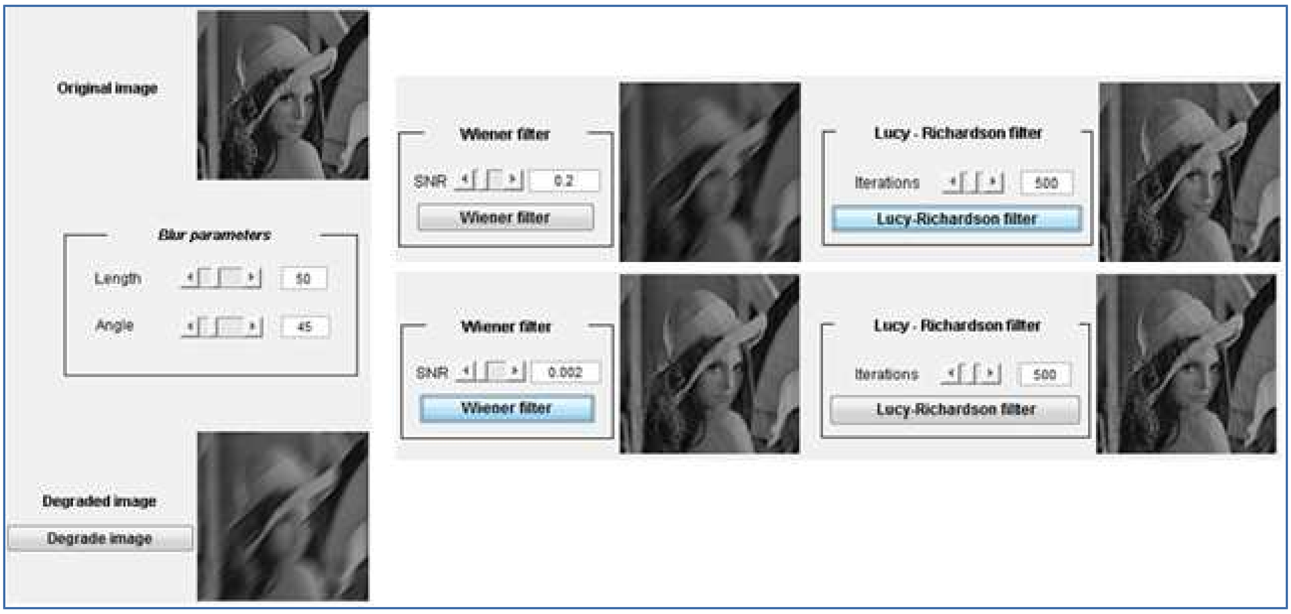
SNR 0.2

Iterations 50

SNR 0.002

Iterations 50

Fig. 2 First experimental results



Length 50

Angle 45

SNR 0.2

Iterations 500

SNR 0.002

Iterations 500

Fig. 3 Second experimental results

In the third experiment, blur parameters leave the same, SNR value is 0,2, and iterations for L-R filter are 500. Restoration with L-R filter in this test is so successful i.e. restored image is comparable with original image. If we change only SNR value to 0.002 and compare the restored image by Wiener filter with restored image by L-R for 500 iterations, we can notice that this two restored images are closest with the original image. These results visually are presented on figure 3. If we compare the restored images with Wiener filter and Lucy - Richardson filter in all experimental cases, generally better results are obtained with Lucy - Richardson filter. This is also shown in the results obtained for image quality measurements, which are calculated directly in the created GUI. If they are analyzed, it can be concluded that the LR filter provides values for image quality measurements that are closer to the ideal values, in comparison to the Wiener filter. This can be seen from the figure that follows, where the data for image quality measurements for 50 Length, Angle 45, SNR 0,2 and LR iterations 50th are presented.

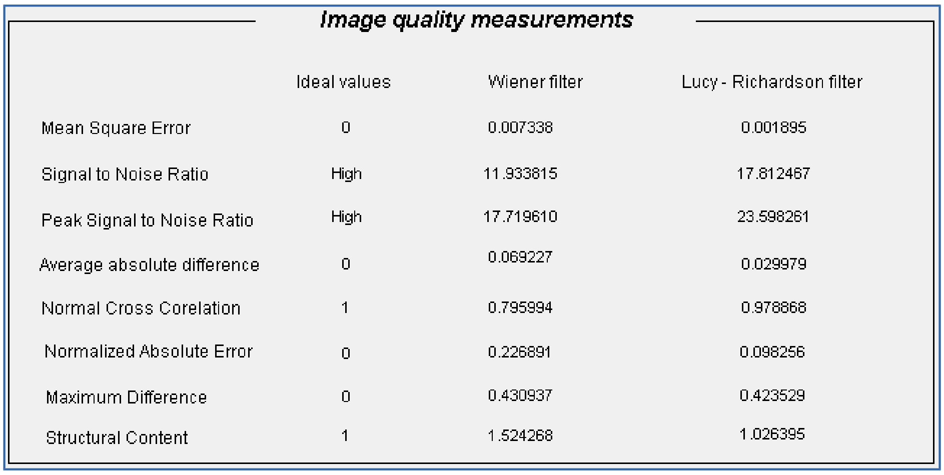


Image quality measurements

	Ideal values	Wiener filter	Lucy - Richardson filter
Mean Square Error	0	0.007338	0.001895
Signal to Noise Ratio	High	11.933815	17.812467
Peak Signal to Noise Ratio	High	17.719610	23.598261
Average absolute difference	0	0.069227	0.029979
Normal Cross Corelation	1	0.795994	0.978868
Normalized Absolute Error	0	0.226891	0.098256
Maximum Difference	0	0.430937	0.423529
Structural Content	1	1.524268	1.026395

Fig. 4 Image quality measurements

VI. CONCLUSIONS

This paper aims to present a graphical user interface - GUI for image restoration and results obtained with it. The graphical user interface was developed in MATLAB 7.10.0 (R 2010a) and has three functions. Firstly, it degrades the original image loaded by adding some blur to it. This can be performed by setting the basic parameters of the blur: length and angle. The second functionality is to restore the degraded image. We have decided to conduct the restoration by using the Wiener and Lucy - Richardson filter, which are the two most commonly used filters for this purpose. Taking the fact into consideration that the Wiener filter is an example of a linear filter and Lucy - Richardson filter is an example of a nonlinear iterative filter, the graphical interface can be used to compare the two types of filters, by comparing the results obtained from both. The third function is related to the analysis of image quality measurements. The analysis in this graphical user interface for image restoration is done through a graph of eight parameters for image quality. This survey shows that for smaller values of the signal - noise ratio (SNR), better results are achieved with the Winner filter. The higher the ratio, the restored image is far from the original.When the number of iterations is larger in Lucy – Richardson filter, the restored image is closer to the original. But by increasing the number of iterations the restoration process is slower, which can be considered as deficiency. Comparing the overall results obtained by the two filters, Lucy-Richardson filter can be used to achieve better results.

REFERENCES

[1] Gerard Blanchet, M. C. (2006). Digital Signal and Image Processing Using MATLAB. London, UK: Hermes Science Europe Ltd.

[2] Jain, "Fundamentals of Digital Image Processing", Engelwood Cliff, N. J.: Print ice Hall, (2006).

[3] Bovik, A. (2009). The essential guide to image processing. London: Academic Press is an imprint of Elsevier. Gerard Blanchet, M. C. (2006). Digital Signal and Image Processing Using MATLAB. London, UK: Hermes Science Europe Ltd.

[4] S. K. Satpathy, S. Panda, K. K. Nagwanshi and C. Ardil, “Image Restoration in Non Linear Filtering Domain using MDB Approch”, International Journal of Signal Processing 6:1 (2010).

[5] A. M. Eskicioglu and P. S. Fisher, “Image quality measures” *IEEE Trans. Communications*, vol. 43, pp. 2959–2965, Dec. 1995.

Poetry as a Visual Novel – a Multimedia Project

Katerina Bashova, Veno Pachovski
School of Computer Science and Information technology
UACS
Skopje, Republic of Macedonia
iris.rice.1998@gmail.com, pachovski@uacs.edu.mk

Abstract— **The development of IT and WWW i.e. Internet poses new challenges. Having that in mind, the prevailing content most sought after is multimedia and the requirements for it grow stronger in time. Poetry is considered one of the oldest forms of literary art. To write poetry requires skill and special command of language. Reading poetry can be a simple pleasure or leisure, but in order to fully understand the message or meaning of a poem, special attention and concentration is needed. If poem is read by an actor or an experienced speaker, and if the voice is augmented by music background, followed by some visual effects, it becomes a real multimedia treat. While reading poetry can be a real pleasure or a deep spiritual experience, to reach that kind of complexity in a multimedia project, specialized software is required. Usually, poetry is organized as a collection of short texts. Its characteristics (short form, human voice, music, rather simple visual illustrations) make it ideal candidate for multimedia project. Nevertheless, transferring a collected poetry into a multimedia project can be a challenging task. Visual novel satisfies those requirements. This paper represents such an attempt.** (*Abstract*)

Keywords— multimedia, poetry, visual novel, Ren'Py (key words)

I. Introduction

A. Visual poetry

The origins of the visual novel date from the 1980s with the game The Portopia Serial Murder Case (jap. ポートピア連続殺人事件 or Portopia Renzoku Satsujin Jiken), which was first published in the Japanese market in 1983 by the company Enix (today's Square Enix). The genre of this game was adventure and it was made for the Japanese computer NEC PC-6001. Later in 1985, the company Chunsoft adapted the game for NES (Nintendo Entertainment System) and this version of the game has become precursor of the visual novels. [1][2][3]

However, the company Chunsoft published the first novel in 1992 and it was the horror thriller game Otogirisō (jap. 弟切草, eng. St John's wort) which was made for Super Nintendo Entertainment System. This game, with 300 000 copies sold, was so successful that it became a role model for the new genre and the games where named as Sound Novels (jap. サウンドノベル or saundo noberu). One specific characteristic of these games is that after the first play, with new start and new decisions made, a new additional scenario would be unlocked. The company Leaf, which was one of the many that wanted to develop this type of games, in order not to use Chunsoft's trademark Sound Novel, named their series games Visual Novels. Leaf's first games were not so successful until 1997, when they published the game To Heart (jap. トゥハート or Tu Hāto). With its warm and touching love story, as well as its high quality music, it became instant success. This motivated other companies to accept this term and since then these types of games are known as Visual Novels. [4]

II. Definitions

A. Visual Novel

The term visual novel (jap. ビジュアルノベル or bijuaru noberu) represents a multimedia game which has all the multimedia's elements like text, backgrounds, characters, music, sounds and it has interaction with the player. According to Cavallaro (2010:8) "the visual novel typically articulates it's narrative by means of extensive text conversations complemented by lovingly depicted (and mainly stationary) generic backgrounds and dialogue boxes with character sprites[1] determining the speaker superimposed upon them".

Typical for the visual novel is the branching of the story, because there are decision points in the game where the player is expected to make a decision about the next step of the story. Based on the chosen option, the player will reach a certain end of the story. This way, when the player finishes the game, he/she is motivated to play it again in order to see the other (alternative) ends of the story. Based on how many decision points are in the visual novel, it increases its alternative story's ends as well as its complexity.

If the given story has only one end and not so much interaction with the player, the visual novel is narrative and linear and is called kinetic novel.

Another specific element of the visual novel are the backgrounds. According to Cavallaro (2010:8) "at certain pivotal moments in the story, more detailed images drawn especially for those scenes and enhanced by more cinematic

[1] character sprites are two-dimensional images

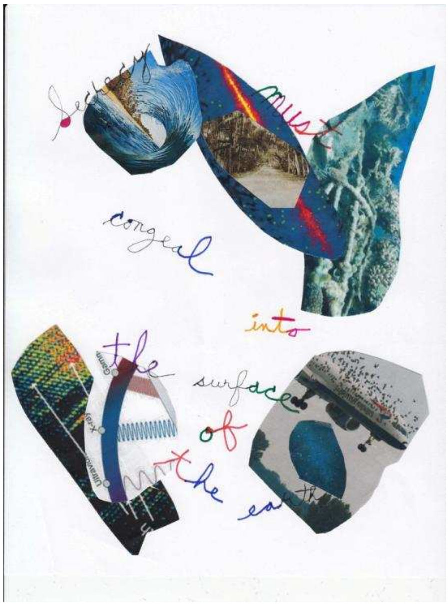
camera angles and CGI[2] are included" to gain the attention and to move the player.

Beside the dialogue, backgrounds and characters, the visual novels include background music and sounds, especially composed and made for the game to capture the story's atmosphere.

The dialogue box in which all the character's dialogue goes, can be a rectangle frame located at the end of the screen or it can be a full screen square-like frame where the text will overlay the background and the characters that appear in the scene. But regardless of the size and shape of the dialogue box, the player can always hide it for a brief moment in order to see the whole scene and admire it.[5]

B. *Visual poetry*

Because of the uniqueness of the form of this visual novel that was developed, it can be considered as some form of visual poetry. In this paper, the subject of visual poetry will not be the studied too deeply.



Source: Nil Almost by Joel Chace, www.archive.org

Fig. 1 Secrecy must congeal into the surface of the earth

Suffices to say that the visual poetry can be considered as poetry or art, hence it is contracted by multimedia elements, like picture, text, audio (combination of sounds, music and narration), and video and so on. According to Klaus Peter Dencker (2000), "visual poetry is the changeable relationship of visual art and literature, of picture and text, of figurative and semantic elements, the connection of both art forms in an intermedial space, the sensory reaction to any kind of communications form coming from the environment, the reservoir for important recognitions from collage, concept art, concrete art, used by different imaginative varieties of realism, for the establishment of evidence and in all the conceivable ways that a logical language deploys." [6] In the new 21st century, where the digital era reigns, the visual poetry can be interactive and its verses can appear on the screen with motion and transformation by a simple click or tab by the reader. Finally, visual poetry can be seen as a possible form of expression in the development of our information- and communications society. Visual poetry can react to the new forms of media (video, computer, holography, laser, and so on), is a form of expression independent of a certain medium, which can enter creatively and innovatively into interactive communications models.

III. The Ren'Py engine

Ren'Py is a visual novel engine that provides easy implementation of all multimedia elements with its special Ren'Py script language. According to Ren'Py's founder Tom Py, "the Ren'Py's script language, which is easy to learn, provides effective writing of huge visual novels, while its Python script allows creating more complex stimulation games. The engine is open source, which means that it can be downloaded freely from Ren'Py's official website, as well its code to be modified and upgraded. Tom Py named the engine Ren'Py, using the words "Ren'ai (from the Japanese term for ren'ai shimyurēshon gēmu - romantic stimulation game or dating sims) and Python, because Ren'Py is developed with Pygame and the programming language Python." [8]

Ren'Py's first official version was released on August 24th, 2004 and it was Windows compatible only. Later. on September 8, 2005 a Macintosh version was released, while the Linux version was released on April 30, 2006. All this versions provided development and release of games for all three platforms. On February 7, 2011, a version was released with support for the special RAPT – the Ren'Py Android Packaging Tool. With this tool, the Ren'Py games could be adapted to be executable on Android tablets and smart phones. [9] [10]

The latest version is Ren'Py 6.17 codenamed "In This Decade…" and it was released on February 20th, 2014. The newest version has these features:

- A rewrite of the Style system that should improve Ren'Py's performance.
- A new style statement that makes it easier to define styles.
- A rewritten shift+I style inspector lets you view those styles.
- A new "show layer" statement that makes it convenient to apply transforms and ATL transforms to entire layers at once.
- A new "window auto" statement that enables automatic management of the dialogue window.
- Several other syntax improvements.
- French and Russian translations.
- The integration of RAPT (the Ren'Py Android Packaging Tool) with the Ren'Py SDK. Ren'Py now

[2] CGI stands for Computer-Generated Imagery

downloads RAPT using the Ren'Py updater - it's no longer necessary to download RAPT separately. [11]

It provides developing games with complex translation system that will allow the player to easily change the languages, and provide the programmer with clarity of the translation as well as options for adding new languages. Additionally, it has improved Android support, including the ability to build APKs from the launcher, support for Expansion APKs, and support for television-based consoles like the OUYA. With its multiplatform support, dedicated team that has been upgrading it actively for nearly nine years, as well as its dedicated forum, the Ren’Py engine is becoming more and more popular and the beginners as well as the companies have created many noncommercial and commercial visual novels. [10][11]

Fig. 2. Ren’Py’s logo [8] (top left), Pygame’s logo[13] (top right) Logo for the novel on Android[14] (down)

With its multiplatform support, dedicated team that has been upgrading it actively for nearly nine years, as well as its dedicated forum, the Ren’Py engine is becoming more and more popular and the beginners as well as the companies have created many noncommercial and commercial visual novels.

IV. Development of the visual novel

For creation of a standard visual novel, it is required to have a story as the base from which the character’s dialogs will be extracted and, based on them, a scenario will be constructed.

Since this is a special case of visual novel, the requirements were completely different. A collection of short poems was chosen, and there was a recording (recorded by a tape-recorder) from a live radio show in which some of them were read by a professional speaker, followed by background music.

All poems were separately integrated into the game, so that they can easily be modified in the future. Because there are six parts in the collection and each part has different number of poems, a list of the parts was made in the main script.py which pointed out to six different files, corresponding to the parts. Those files contain a menu list of the locally stored poems.

At the same time, background music was chosen, buttons and interface were designed using Adobe Photoshop© and editing of the recording was done with Adobe Soundbooth©. Every button was carefully edited to correspond to the five states defined in Ren’Py: ground or base, idle, selected, selected_hover and selected_idle. For the save and load screen, a special interface was designed - a screenshot of a save/load point with short description is placed on a scrub with wax stamp.

The background music wasn’t edited, but the narration was heavily edited due to its age (1993) and corrupted recording. Before the editing was started, in the root game folder many sub folders were created, illustrated in the figure below.

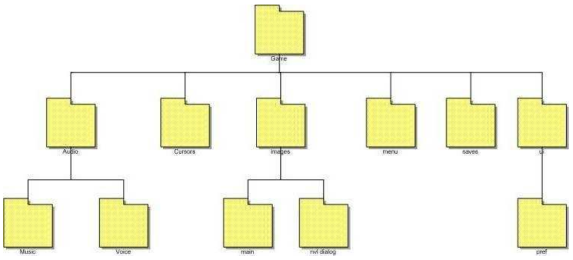


Fig. 3. Diagram of tree structure of the root game folder

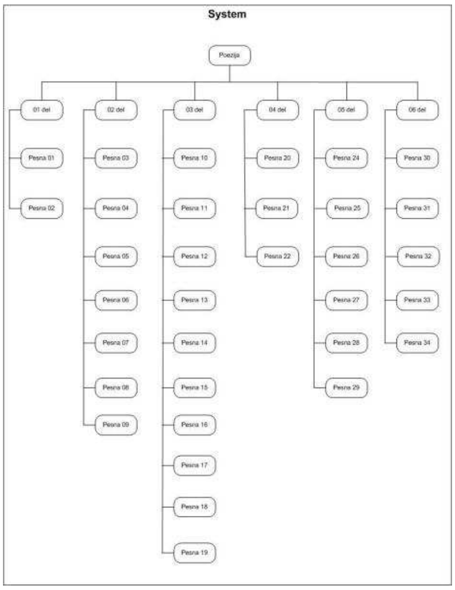


Fig. 4. Tree dependency of the poems

Then every poem was extracted from the recording, each edited separately and saved into the subfolder Voice. Similar to the narration, the music was stored into the Music folders, the images in to the separate subfolders of the Images folder and so on.

After all the graphical and audio elements were placed in there corresponding folders, the editing of the poems was started using Ren’Py and they were connected with the corresponding audio elements. At the same time, the interface of the game was coded.

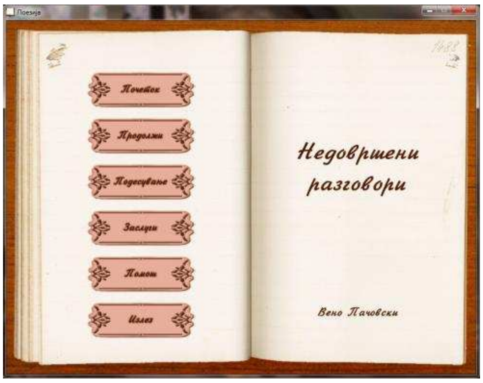
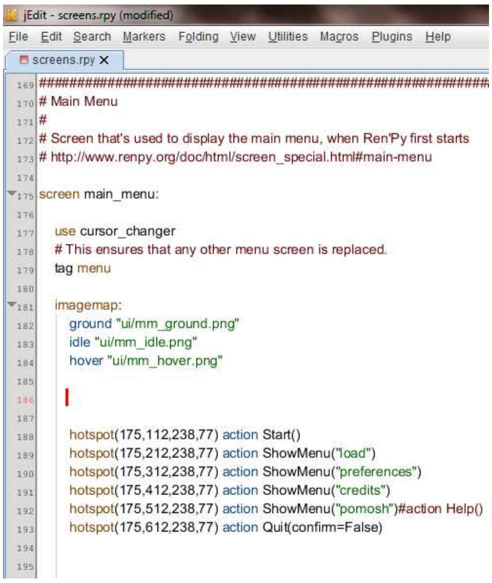


Fig. 5. Main menu


jEdit - screens.rpy (modified)
File Edit Search Markers Folding View Utilities Macros Plugins Help
screens.rpy

```
##############################################################
# Main Menu
#
# Screen that's used to display the main menu, when Ren'Py first starts
# http://www.renpy.org/doc/html/screen_special.html#main-menu

screen main_menu:

    use cursor_changer
    # This ensures that any other menu screen is replaced.
    tag menu

    imagemap:
        ground "ui/mm_ground.png"
        idle "ui/mm_idle.png"
        hover "ui/mm_hover.png"



        hotspot(175,112,238,77) action Start()
        hotspot(175,212,238,77) action ShowMenu("load")
        hotspot(175,312,238,77) action ShowMenu("preferences")
        hotspot(175,412,238,77) action ShowMenu("credits")
        hotspot(175,512,238,77) action ShowMenu("pomosh")#action Help()
        hotspot(175,612,238,77) action Quit(confirm=False)
```

Fig. 6. Code of Main menu in Ren'Py

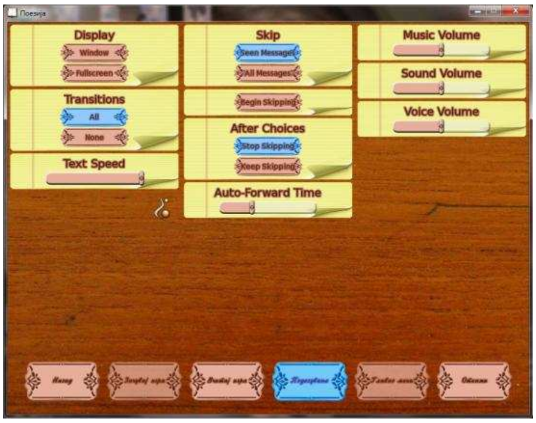
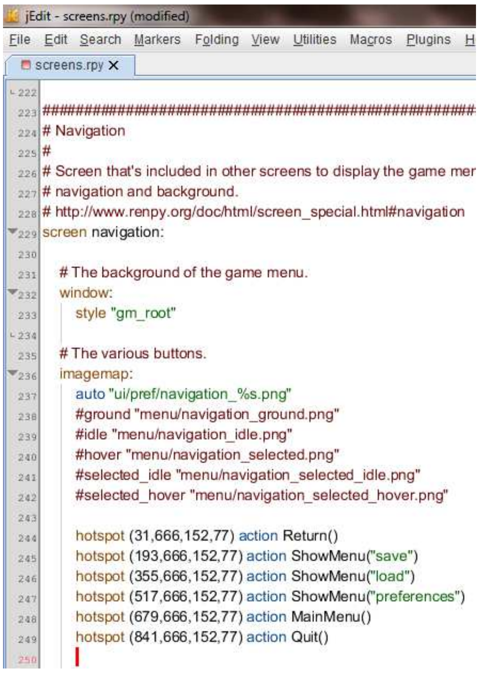


Fig. 7. The Preferences menu


jEdit - screens.rpy (modified)
File Edit Search Markers Folding View Utilities Macros Plugins H
screens.rpy

```

##############################################################
# Navigation
#
# Screen that's included in other screens to display the game mer
# navigation and background.
# http://www.renpy.org/doc/html/screen_special.html#navigation
screen navigation:

    # The background of the game menu.
    window:
        style "gm_root"

    # The various buttons.
    imagemap:
        auto "ui/pref/navigation_%s.png"
        #ground "menu/navigation_ground.png"
        #idle "menu/navigation_idle.png"
        #hover "menu/navigation_selected.png"
        #selected_idle "menu/navigation_selected_idle.png"
        #selected_hover "menu/navigation_selected_hover.png"

        hotspot (31,666,152,77) action Return()
        hotspot (193,666,152,77) action ShowMenu("save")
        hotspot (355,666,152,77) action ShowMenu("load")
        hotspot (517,666,152,77) action ShowMenu("preferences")
        hotspot (679,666,152,77) action MainMenu()
        hotspot (841,666,152,77) action Quit()
```

Fig. 8. Code for the navigation in the Preferences screen

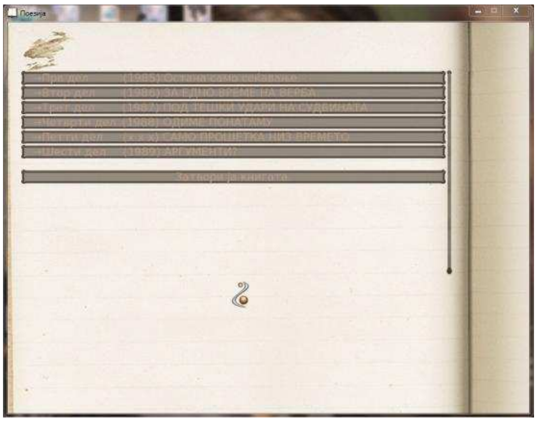

Поезија
→Прв дел (1985) Остана само сеќавање
→Втор дел (1986) ЗА ЕДНО ВРЕМЕ НА ВЕРБА
→Трет дел (1987) ПОД ТЕШКИ УДАРИ НА СУДБИНАТА
→Четврти дел (1988) ОДИМЕ ПОНАТАМУ
→Петти дел (x x x) САМО ПРОШЕТКА НИЗ ВРЕМЕТО
→Шести дел (1989) АРГУМЕНТИ?
Затвори ја книгата

Fig. 9. The game's main menu containing the six parts

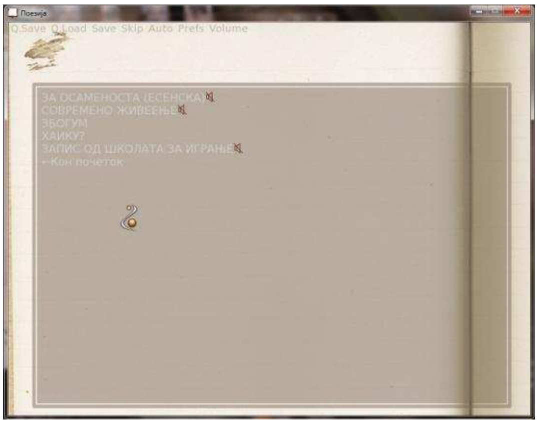
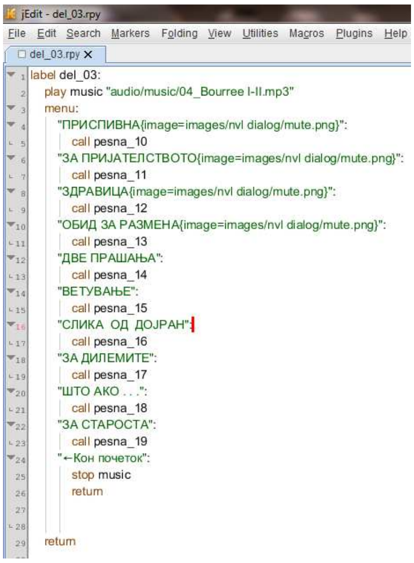


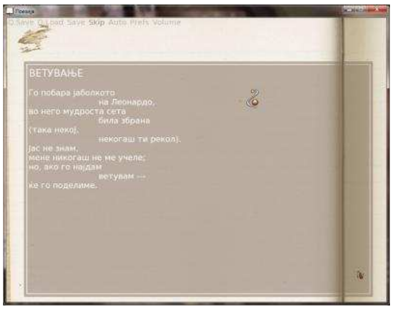
Fig. 10. The contents of one part of the collection


jEdit - del_03.rpy
File Edit Search Markers Folding View Utilities Macros Plugins Help
del_03.rpy

```
label del_03:
  play music "audio/music/04_Bourree I-II.mp3"
  menu:
    "ПРИСПИВНА{image=images/nvl dialog/mute.png}":
      call pesna_10
    "ЗА ПРИЈАТЕЛСТВОТО{image=images/nvl dialog/mute.png}":
      call pesna_11
    "ЗДРАВИЦА{image=images/nvl dialog/mute.png}":
      call pesna_12
    "ОБИД ЗА РАЗМЕНА{image=images/nvl dialog/mute.png}":
      call pesna_13
    "ДВЕ ПРАШАЊА":
      call pesna_14
    "ВЕТУВАЊЕ":
      call pesna_15
    "СЛИКА  ОД  ДОЈРАН":
      call pesna_16
    "ЗА ДИЛЕМИТЕ":
      call pesna_17
    "ШТО АКО . . .":
      call pesna_18
    "ЗА СТАРОСТА":
      call pesna_19
    "←Кон почеток":
      stop music
      return


  return
```

Fig. 11. Code for part 3 of the collection, containing the links to files with the poems



ВЕТУВАЊЕ

Го побара јаболкото
на Леонардо,
во него мудроста сета
била збрана
(така некој,
некогаш ти рекол).
Јас не знам,
мене никогаш не ме учеле;
но, ако го најдам
ветувам —
ќе го поделиме.

Fig. 12. Structure of a poem with narration

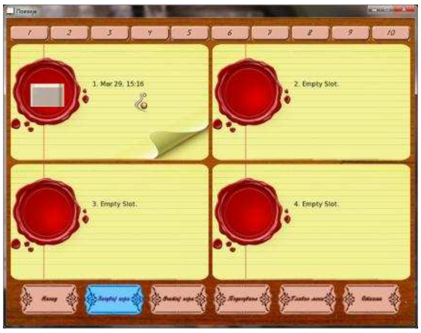


Fig. 13. The Save screen

V. CONCLUSION

Based on the researched, it can be concluded that the visual novel can be used as a way to present the visual poetry. The branching in the visual novel, which is the main characteristic, can be used to implement multiple poems by multiple poets that can be interactive, have music, sounds and backgrounds which will present the poems.

The Visual novel can be a challenge for the programmers, because the development consists of coding in two languages, Ren'Py and Python, designing the backgrounds, the characters, requires composing the background music, recording sounds and character's voices as well as defining coding strategy based on the story's scenario. Hence, a team of programmers, background's and character's artists, authors, composers and voice actors is needed to power the visual novel's development.

The visual novel's development, depending of its complexity and the team's size, can take from half year to year and half, if the whole project is well managed and the team well coordinated.

REFERENCES

[1.] Portopia Renzoku Satsujin Jiken [accessed, 31 December 2013] http://en.wikipedia.org/wiki/The_Portopia_Serial_Murder_Case

[2.] John Szczepaniak 'Portopia Renzoku Satsujin Jiken', "Retro Gamer" issue 85, 2011

[3.] Visual Novel Database, 'Portopia Renzoku Satsujin Jiken', added: 29 June 2010, [accessed, 31 December 2013]

http://vndb.org/v4511

[4.] 'Visual novel'

[accessed, 31 December 2013]
http://en.wikipedia.org/wiki/Visual_novel
[accessed, 4 April 2013]

[5.] Cavallaro, Dani "Anime and the Visual Novel: Narrative Structure, Design and Play at the Crossroads of Animation and Computer Games", Jefferson, McFarland & Company, 2010 [Online]

[accessed, 31 December 2013]
http://mcfarlandpub.com/book-2.php?id=978-0-7864-4427-4
[Accessed, 31 December 2013]

[6.] Dencker, Klaus Peter "From Concrete to Visual Poetry, with a Glance into the Electronic Future", Kaldron On-Line and Light and Dust Mobile Anthology of Poetry, 2000

[accessed 30 March 2014]
http://www.thing.net/~grist/l&d/dencker/denckere.htm

[7.] Chace, Joel "Nil Almost", Shuf Poetry; Truck; The Jivin' Ladybug Poema Visual, August 7, 2013

[accessed 30 March 2014]
http://www.thing.net/~grist/l&d/dencker/denckere.htm

[8.] Ren'Py's official website [Online]

[accessed, 31 December 2013]
http://www.renpy.org/

[9.] Why Ren'Py?'

[accessed, 31 December 2013]
http://www.renpy.org/why.html

[10.] 'Ren'Py Release List'

[accessed, 31 December 2013]
http://www.renpy.org/release_list.html

[11.] 'Download Ren'Py 6.17'
[accessed 30 March 2014]

http://www.renpy.org/latest.html

[12.] Download Ren'Py 6.16'
[accessed 30 March 2014]

http://www.renpy.org/release/6.16

[13.] Pygame's official website

[accessed, 31 December 2013]
http://www.pygame.org/news.html

[14.] 'Download Ren'Py 6.12.0'

[accessed, 30 March 2014]

Sampling Method for Size Reduction of Stochastic Model for Peer-assisted VoD streaming

Sasho Gramatikov
Faculty of Computer Science and Engineering, Skopje,
University ”Ss. Cyril and Methodius”, Skopje, Macedonia
Email: sasho.gramatikov@finki.ukim.mk

***Abstract*—Video on Demand is a leading TV service offered by the IPTV operators in the past decade that has been rapidly gaining on popularity because it offers great convenience to the customers to watch any video they want at any time. However, the drawback of this service it that it is very resource demanding and expensive for the operator. One of the solutions for reducing the high traffic demands from the video servers is the implementation of peer-assisted streaming, i.e., including the peers in the streaming process by taking advantage of their unused streaming and storage capacity. In order to estimate the reduction of the traffic demand depending on the network configuration and the intensity of demands for videos, we developed a stochastic model that determines the system behavior in stationary state. We proved that the model is a precise tool for estimating the server demands. However, the model requires high computation power for obtaining the desired results. The size of the number of linear equations that have to be solved grows rapidly with the growth of the number of peers and their streaming capacity, which is serious issue for using the model as a tool for estimation of the system performance. Therefore, we propose a sampling method that significantly reduces the size of the system of linear equations, and thus, reduces the computation time and resources required for obtaining the results. Our analysis shows that although the size of the system is reduced, the relative error compared to the original unreduced system is negligible.**

***Keywords*—VoD, Peer-assisted streaming, stochastic model, size reduction**

I. Introduction

The high popularity of the video contents made a solid ground for the video domination in the global consumer traffic. The videos are becoming so popular that, according to the estimations, by the year 2017, they will occupy 80-90% of the globally exchanged traffic worldwide [1]. A significant part of this traffic will belong to the Video on Demand service (VoD) which will triple the amount of traffic that it generates nowadays. The main reason for its popularity is that it gives the clients the freedom to watch a large variety of videos at any time. Moreover, it has become more accessible to a larger community as a result of the expansion of the Internet Protocol Television (IPTV). However, the delivery of these traffic-intensive contents requires a separate unicast flow for every request of the clients, which is a serious burden for the delivery network and the streaming servers. In addition, the demand for higher quality videos and the growing popularity of the service further increases the amount of the traffic which threatens to congest the delivery networks. Therefore, finding an optimal way to deliver the videos in the network has been a challenging task for the network operators and the research community.

One of the approaches used to solve the problem of network congestion is the Peer-to-Peer (P2P) concept. As it proved to be a successful solution for sharing large files in the Internet, this concept has encouraged many researchers to consider its implementation in live and VoD streaming in the Internet [2], [3]. However, the main issue of the P2P for delivery of video contents is the real-time nature of the VoD service, i.e., the blocks of video data have to be provided in consequent order, at the moment when they are requested. Unlike the pure P2P systems, a more common practice in the managed environments is the use of the peer-assisted streaming where the peers take an important role in the streaming. The peers help to reduce the traffic in the network core, but the servers still remain an inevitable part of the system that guarantees the required QoS. These solutions together with various content distribution schemes prove to be efficient in reducing [4], [5], [6], [7], [8], [9] and even eliminating the traffic requested from the streaming servers [10], [11].

In our work, we consider wall-gardened networks owned and managed by the IPTV service provider because the network can be completely controlled by the operator and configured according to the offered services and user demands. We focus on the concept of peer-assisted streaming where the peers will help to reduce the traffic in the core of the wall-gardened network, but the servers will still remain an inevitable part of the system that will guarantee the required quality of service [4]. Apart from the design of a system for peer-assisted streaming, our main goal is to mathematically model its behavior and analyze how some environmental parameters influence the performance. Therefore in our previous work [12], we proposed a system for peer-assisted VoD contents in managed networks. In that work, we focused on the design of a mathematical model for the proposed system and used it to analyze how some environmental parameters influence the performance. The system is represented as a closed network of queues with finite customers. Depending on the number of customers served by these queues in a given moment of time, we defined a state diagram where each node is the probability of the given state. These probabilities are interconnected with arc that represent the probability flows of the probabilities.

The main issue of this kind of representation of the system is that the number of states significantly increases with the increasing number of clients in the network and their streaming capacity. The size of the diagram is so big, that the values of the probabilities of the states cannot be calculated with a Personal Computer (PC) even for networks of few hundreds of peers. This is a concerning fact because the model, although precise, cannot be used for estimating the performance of the peer-assisted streaming in networks with higher scale. Therefore, we propose a sampling method for reduction of the size of the state diagram without significant errors in the estimation. The main idea behind the sampling process is choosing states with equal constant distance between them and discarding the remaining states of the diagram. In the process of sampling, we reconnect the arcs that lead to the discarded states or terminate in the chosen states from the discarded states. Thus, we obtain a significantly reduced size of the system that can be easily computed even for large networks.

The rest of the paper is organized as follows: After the description of the system for peer-assisted streaming and its stochastic representation in II, in Section III, we present the sampling method for reduction of the size of the state diagram of the model. In Section IV we compare the the introduced releative error due to the size reduction for various values of the sampling step. Eventually, we give the conclusions from our work in Section VI.

II. Mathematical model overview

In our work in [12], we proposed a stochastic model for modeling a system for peer-assisted VoD streaming. The system is designed for managed networks where the operator has the control over the clients' resources, and therefore, can take advantage of their unused storage and streaming capacity. The operator pre-populates the STBs of the clients with strips of the videos in the off-peak hours according to different distribution schemes. In the system, the geographically close clients are grouped into local communities, which are served by a dedicated streaming server. Apart from streaming videos to the clients, the streaming server also has a role of an index server for its local community. It maintains availability data of the contents on the clients, as well as data related to the occupancy of the streaming capacity of the peers. Upon reception of a request for a video, the streaming server searches for peers that have available copy of the video and enough streaming capacity and assigns these peers to serve the client. If one video cannot be entirely served by the peers because parts of the video are not stored in the peers or because they do not have available streaming resources, the rest of the video is served by the streaming servers. Although the clients participate in the streaming, the streaming servers still have the main role of streaming the videos, while the clients only alleviate their load.

The video content library contains c videos with playback rate r. For modeling the users' behavior not to watch the entire video, it is considered that the average duration of the videos d is actually the average time they spend watching the videos. The items in the library are sorted according to their popularity, beginning with the most popular video. Because of the limited streaming capacity of the clients and their inability to stream entire video uninterruptedly, the videos are divided into m parallel strips with equal size. Thus, each peer can assist in the streaming of a video at least with one strip. Instead of storing the entire videos in the peers, they are first divided into strips, and then, the strips are independently distribute in the peers. This approach increases the availability of the contents on the peers and the utilization of their uplink capacity. The time necessary to stream one strip is equal to the average watched video duration d, however, the streaming occupies m times less uplink capacity. The rate r/m necessary to stream one strip is defined as a channel. The probability that a video content with rank v will be requested by a peer is marked as $P(v)$.

The number of active peers in a local community is denoted as n. This is the number of the peers that are connected to the network and have their STB switched on. The peers in the model are reliable, i.e., they never turn the STB off. Each peer has an STB that has capacity to store s strips of video contents. The downlink capacity of the STB is higher than the playback rate of the content items, while the uplink capacity is expressed as the number of simultaneous channels k that the peer is capable to stream.

In the modeling of the proposed system, the entire process of serving the clients is presented as a stochastic process which is brought to a closed analytical form using the foundations of the queuing theory [13] and the generating functions [14]. The model takes as input the size of the local community, the storage and streaming capacities of the peers, the size of the videos, the intensity of requests, the size of the video library, and what is most important, the distribution of the contents in the peer. The output of the mathematical model is the utilization of the streaming capacity of the peers, which is further used for calculation of the peer-assisted traffic in the system and the traffic served by the servers.

The process of generating requests for videos by the clients is a Poisson process with arrival rate λ, i.e., the inter-arrival time $w = 1/\lambda$ of two consequent requests has an exponential distribution. Since each request for a video is a request for m parallel strips, the serving of the requests by the server or by the peers is also a Poisson process with service rate $\mu = 1/d$.

Since both the request and serving process are Poisson processes, and the number of clients in the network is constant, the entire system is represented as a closed network of queues with finite customers. The number of available customers is mn, which is the total number of possible requests for strips. In a given moment of time, the customers can be either waiting to make a request for a strip or are being served by the peers or the server. The network of queues is shown in Fig. 1

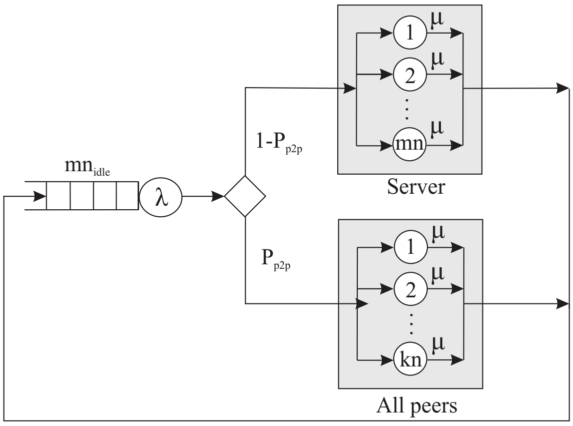


Fig. 1. Representation of the system for peer-assisted VoD streaming with closed network of queues with finite customers

The network of queues in Fig. 1 consists of a waiting queue and video receiving queues. The peers are initially in the waiting queue, where they stay until they make a request. The average size of the waiting queue in stationary state is mn_{idle}. When a peer makes a request for a strip, it can be served either by other peers or by the server. Therefore there are two receiving queues where a request enters depending on the availability of the requested strip. If there is at least one peer that has a copy of the requested strip and an available channel, it will be served by the peers. The probability of such an event is P_{p2p}, which is a function of the contents' availability and the current state of the system. Since the system has n peers that can simultaneously stream k parallel strips, the streaming of the requests by the peers is modeled as a queue with nk parallel service facilities $M/M/nk$. If the requested strip is not available on the peers or it is available, but none of the peers that contains the strip has available channels, the request is served by the server with probability $1 - P_{p2p}$. Since the server has capacity to serve all those strips that the peers cannot serve, the server is modeled as a queue with mn parallel service facilities, i.e., $M/M/mn$ queue. After the end of the streaming, the peers leave the video receiving queues and re-enter the waiting queue. The main task of the mathematical model is to find the average size in stationary state of all these queues, so that the portion of the traffic served by the peers can be determined.

In order to fulfill this task we define the stationary state of the system as the number of strips streamed by the peers and number of strips streamed by the server. The probability for arbitrary values i and j, such that $i + j < mn$, is defined as p_{ij}. From the analysis in our previous work [12], we obtain the following dependency of the state probability p_{ij} related to the other state probabilities of the system:

$$\begin{aligned} 0 &= P_{p2p}(i-1)(mn-i-j+1)\lambda p_{i-1,j} + \\ &+ (i+1)\mu p_{i+1,j} + (j+1)\mu p_{i,j+1} - \\ &+ \left(1 - P_{p2p}(i)\right)(mn-i-j+1)\lambda p_{i,j-1} + \\ &- \left((mn-i-j)\lambda + (i+j)\mu\right) p_{i,j} \end{aligned} \tag{1}$$

If we consider the dependencies of all the possible stationary probabilities of the system, we obtain a system of linear equations which, if solved, will give the values of the stationary probabilities of every possible state, and hence, the average number of busy channels in the system. To give a better picture of the states of the system in equilibrium, Fig. 2 shows the state diagram obtained from the system of linear equations Eq. (1).

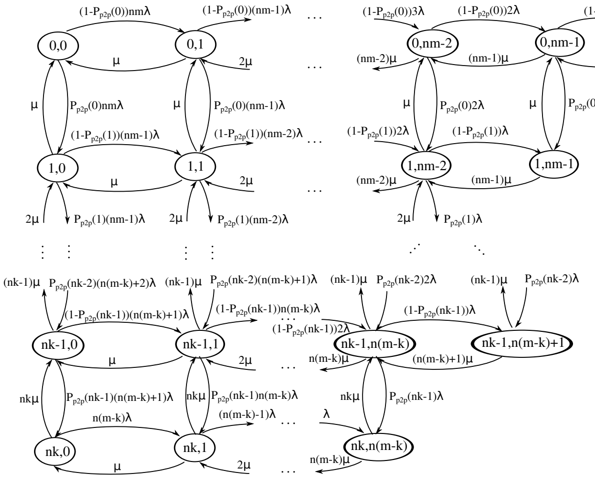


Fig. 2. Diagram of flow of state probabilities.

In the diagram $\mathcal{D}$, shown in Fig. 2, each circle represents the probability of one possible state of the system, and the arrows represent the flow of probability that the system will go from one state to another. For the sake of clarity, the self-loops of the probabilities represented by the coefficient of the last member of Eq. (1) are intentionally omitted. The width of the diagram is always $mn + 1$, while its height depends on the number of streaming channels k of the peers. The first row of the diagram $\mathcal{D}$ has $nm + 1$ states, and each following row has one state less than the previous one. There are $nk + 1$ rows, obtained for each possible number of busy channels on the peers, starting from 0 to nk. By using the expression for the sum of the first n natural numbers $\sum_{i=1}^{n} i = n(n+1)/2$, the expression for the size of the state diagram $\mathcal{D}$ is obtained as:

$$\begin{aligned} \text{size}(\mathcal{D}) &= \sum_{r=0}^{nk}(nm+1-r) = (nk+1)^2 - \sum_{r=0}^{nk}(r) = \\ &= (kn+1)\left(\left(m - \frac{k}{2}\right)n + 1\right) \end{aligned} \tag{2}$$

The system of linear equations Eq. (1) can be easily solved if it is presented in a matrix form:

$$A\mathbf{p} = \mathbf{0} \tag{3}$$

where $\mathbf{p} = [p_{0,0}\ p_{0,1}\ \ldots p_{kn,(m-k)n}]^T$ is the vector of the unknown stationary probabilities of the states of the system with size $((m - k/2)n + 1)(kn + 1)$, and $A = [\mathbf{a}_{0,0}^T\ \mathbf{a}_{0,1}^T\ \ldots\ \mathbf{a}_{kn,(m-k)n}^T]^T$ is the coefficient matrix, where $\mathbf{a}_{i,j}$ is a row vector with the same size as $\mathbf{p}$ such that contains the weights of the arcs on the state diagram in Fig. 2 that go out of the state $p_{i,j}$ on the positions corresponding to the positions of the states in the vector $\mathbf{p}$ that they are directed to. In order to avoid obtaining infinite number of solutions of the system because of the singularity of the matrix A , we use the condition that the sum of all the stationary probabilities equals 1 to modify the matrix A and the vector $\mathbf{0}$. Thus, we obtain a modified matrix B by substituting an arbitrary row in A with row vector $\mathbf{1}^T$ and a modified vector $\mathbf{b}$ by substituting one element in the vector $\mathbf{0}$ with value 1 at the same position as the substituted row in A. With these modifications, the solution of the system is obtained by solving the system:

$$\mathbf{p} = B^{-1}\mathbf{b} \tag{4}$$

With the vector of probabilities $\mathbf{p}$ obtained, the probability that each peers has i busy channels can be obtained by summing all the probabilities of the diagram $\mathcal{D}$ that lay in the same row. Hence the average number of busy channels η in the system can be calculated as the expected number of busy channels. Using this value we calculate the percentage of the overall streaming traffic served by peers as:

$$\theta = \eta \frac{k}{m} \left(1 + \frac{\mu}{\lambda}\right) \tag{5}$$

Although the solution of the system obtained by multiplication of a matrix and a vector Eq. (4) is quite straightforward, the size of the matrices, and hence the number of equations, is a serious issue from a computational point of view. From Eq. (2) we can see that the size of the system rapidly grows with the increment of the size of the community n and the streaming capacity of the peers k. For typical values of the system parameters, e.g., $n = 200$ peers, $m = 10$ strips and $k = 5$ channels, the size of the state diagram will be approximately $1.5 \cdot 10^6$ states. This is a considerably high number when the computation is concerned. Taking into consideration the fact that the size of the coefficient matrix B is determined by the number of states of the system, its size for this specific case of values would be $1.5 \cdot 10^6 \times 1.5 \cdot 10^6$. For a representation of the values of the matrix with double precision numbers that occupy 8 bytes, the overall coefficient matrix B would occupy approximately 11.3 TB of memory space, which is hardly a feasible size of RAM memory with the current computer technology. Moreover, the computation of the inverse B^{-1} and its multiplication with $\mathbf{b}$ in Eq. (4) would also require high computational power. Consequently, the computation of the probabilities of the states would be a very time-consuming and resource-demanding process.

III. Sampling of the state diagram

In order to accelerate the computation and reduce the required resources, we propose a sampling method for reducing the size of the original state diagram $\mathcal{D}$. The sampling consists in forming a new state diagram by choosing each h-th state of the original state diagram shown in Fig. 2, starting from the top-left state and going in both right and down direction. In the process, the states that are not chosen are omitted, however, the probability flows that leave or arrive at these states are included in the new probability flows that connect the chosen states. The sampling of the original state diagram and the connection of the sampled states with the new probability flows is shown in Fig. 3. The new, reduced diagram $\widetilde{\mathcal{D}}$ includes only those states with indexes that are multiple of the sampling step h, provided that the number of channels that can be requested in the system nm and the number of channels on the peers nk are multiples of the sampling step h. The sampling method also includes the right-most and down-most states of the diagram in the case when nm and nk are not divisible by the sampling step h.

A global picture of the reduced state diagram $\widetilde{\mathcal{D}}$ is shown in Fig. 5. Using the same approach for calculating the size of the diagram $\mathcal{D}$ in Eq. (2), the size of the reduced diagram $\widetilde{\mathcal{D}}$ can be presented as:

$$\text{size}(\widetilde{\mathcal{D}}) = \frac{(2J - I + 3)(I + 2)}{2} \tag{6}$$

where

$$I = \lceil nk/h \rceil - 1 \tag{7}$$

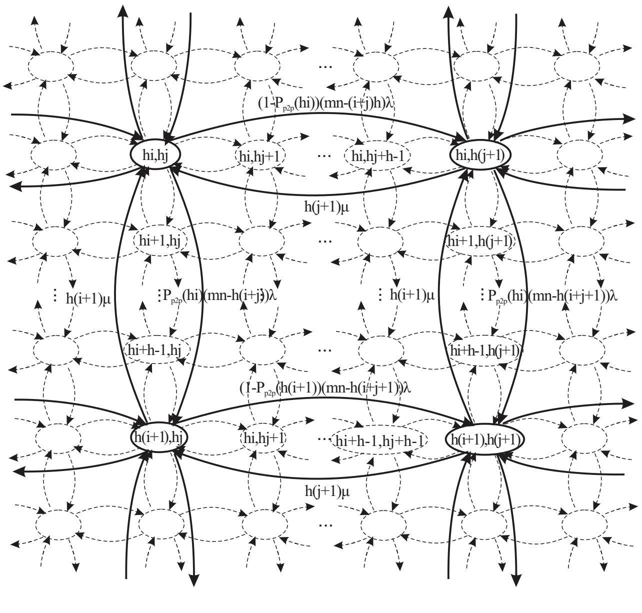


Fig. 3. Sampling of the original diagram of flow of probabilities of a system with cooperative peers.

and

$$J = \lceil nm/h \rceil - 1 \tag{8}$$

In order to present a clearer picture of the sampling process, Fig. 4 shows an example of a small-scale system with $n = 6$ peers, with capacity of $k = 3$ channels, videos divided into $m = 5$ strips and sampling step $h = 7$ states, obtained by substituting the values in the general reduced state diagram shown in Fig. 5. Instead of approximately 300 states obtained from Eq. (2), by using the sampling method, the new size of the diagram $\widetilde{\mathcal{D}}$ according to Eq. (6) is only 18 states.

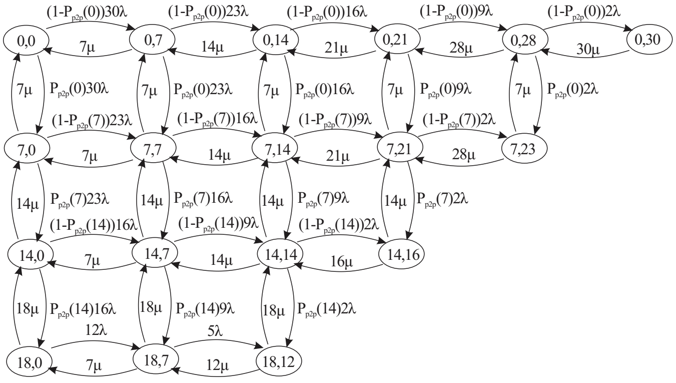


Fig. 4. Example of a reduced diagram of flow of probabilities of a system with cooperative peers.

The system presented in the previous example has very small size, which is not of interest for conducting analysis of real-sized systems. The advantages of the sampling method are more evident in the case of the previously considered system with $n = 200$ peers, $m = 10$ strips and $k = 5$ channels. By choosing a sampling constant with value $h = 20$ states and using it in Eq. (7) and Eq. (8), I and J would be 49 and 99,

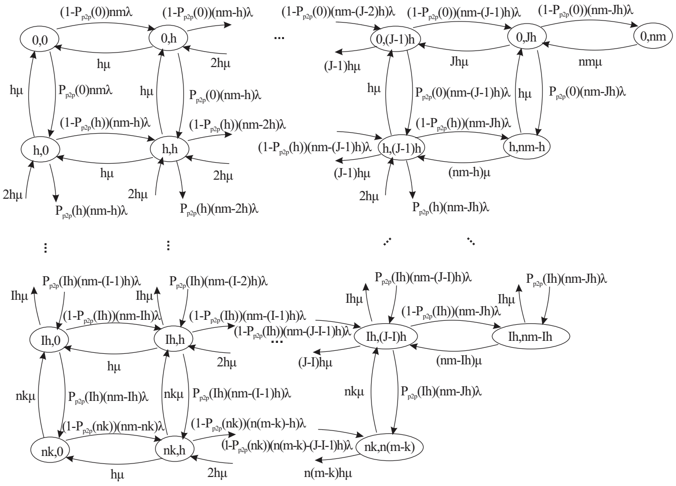


Fig. 5. Reduced diagram of flow of probabilities of a system with cooperative peers.

which substituted in Eq. (6) would give a size of approximately 4000 states. The new reduced diagram has approximately 400 times less states compared to the original diagram with size of $1.5 \cdot 10^6$ states. Increasing the value of the sampling step to $h = 50$ states, would give approximately 650 states, which is equivalent of a gain of 2300 times in the size reduction.

IV. Relative error comparison

Naturally, it is expected that the reduction of the size would introduce errors in the results and that the value of these error would depend on the size of the sampling step h. For that purpose, Fig. 6 shows the dependence of the relative error δ of the peer-assisted traffic θ for a system with size $n = 200$ peers, $m = 10$ strips and storage capacity $s = 100$ strips for variable values of the streaming capacity k and the sampling step h. Since the value of the peer-assisted traffic θ for a system with the given size is hard do be obtained with an ordinary PC, we use the value of obtained for sampling step $h = 2$ as a reference value for calculating the relative errors. Fig. 6 shows that, apart from the sampling step h, the relative error largely depends on the streaming capacity of the peers k. A critical value of the streaming capacity is $k = 8$ when the relative error δ reaches the maximum value. However, it can be concluded that up to $h = 50$ states, the relative error is below 1% for almost the entire range, expect for the critical value $k = 8$ channels, when the relative error has slightly bigger value than 1%.

Choosing the value of the sampling step h will depend on the maximum size of the system that the operator can afford for computation of the results and on the maximum relative error it can tolerate. The exact value has to be a compromise between the resources and the correctness of the results. Tab. I shows an overview of the sizes of the system obtained for various values of the sampling constant h and the streaming capacity k, obtained from Eq. (6). If these results are compared to the relative error in Fig. 6, it can be concluded that choosing a sampling value $h = 20$ considerably reduces the size of the system and at the same time keeps satisfactory level of error in the results for all values of the streaming capacity k.

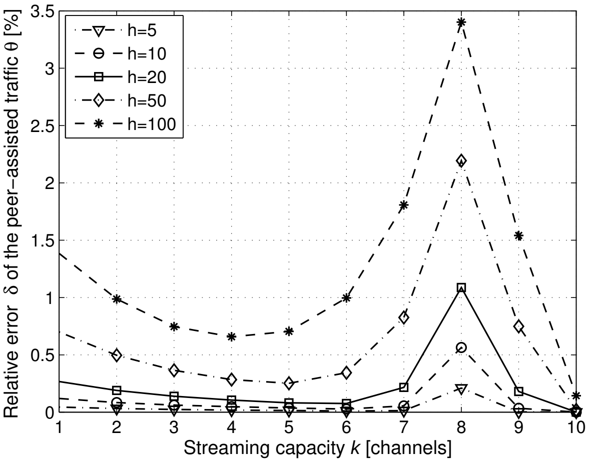


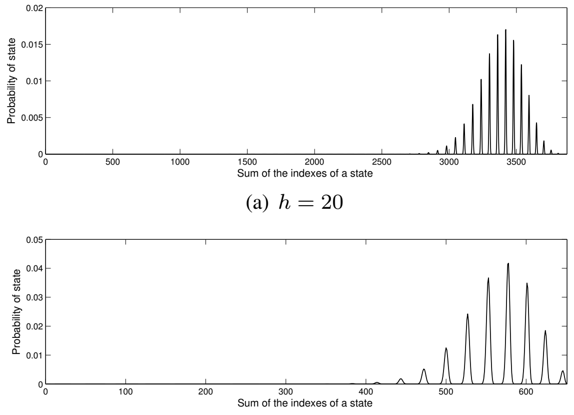
Fig. 6. Dependence of the relative error δ of the peer-assisted traffic θ on the streaming capacity of the peers k and the sampling constant h for community with size $n = 200$ peers and $m = 10$ strips per video.

TABLE I. Overview of the size of the system with $n = 200$ peers and $m = 10$ strips for various values of the sampling step h and the streaming capacity k.

h	Size		
	$k = 2$	$k = 5$	$k = 10$
1	722201	1502501	2003001
2	181101	376251	501501
5	29241	60501	80601
10	7421	15251	20301
20	1911	3876	5151
50	333	651	861
100	95	176	231

V. Dsicussion

One possible solution for a further reduction of the relative error with the same or even smaller size of the system obtained for a certain value of the sampling step h can be proposed if the values of each state probability $p_{i,j}$ shown in Fig. 7 are more thoroughly analyzed. The figure shows two cases of the probabilities of the states of a system with $n = 200$ peers with capacity of $k = 5$ channels and sampling step $h = 20$ and 50 states. The sum of all the values in both the figures equals 1, however, Fig. 7(a) has more states and more bell-shaped curves than Fig. 7(b). An important conclusion that can be made from the figures is that each local bell-shaped curve represents the state probabilities of one row of the diagram in Fig. 5. The bell-shaped curves in the figures presented from left to right are obtained for the rows of the state diagram when moving from the first row, downwards, i.e., the left curves refer to small values of the index i and the right curves refer to large values of the index i. The number of curves is determined by the sampling step in a vertical direction. The movement within a local curve from left to right refers to increasing the value of the index j. The number of values within a curve is determined by the value of the sampling step in horizontal direction. Therefore, the sampling step in vertical direction h_v can have a different value from the step in the horizontal direction h_h. Since the average uplink utilization η, and hence the peer-assisted traffic θ, are obtained by multiplying the value of the index i with the sum of the probabilities of its corresponding local curve, it is more important to have more rows in the diagram. Consequently, the value of the vertical sampling constant h_v can be decreased and the value of the horizontal sampling constant h_h can be increased.



(a) $h = 20$

(b) $h = 50$

Fig. 7. Probability of the states (i, j) of a system with $n = 200$ peers, $m = 10$ strips and $k = 5$ channels for sampling constant (a) $h = 20$ and (b) $h = 50$.

Fig. 7 also shows that for small values of the index i, the local bell-shaped curves are not visible, which means that these states have probability close to value 0, and therefore, can be ignored without introducing noticeable error in the calculations. Therefore, the vertical sampling constant h_v for the small values of i can be increased in order to join the least probable states in the first rows into fewer rows. Although these states will be represented by fewer states, the sum of their probabilities will be again close to value 0, and the increasing of the sampling constant would not introduce errors. On the contrary, for the higher values of the index i, the vertical sampling constant can be reduced so that the probabilities of the more popular states in the last rows of the diagram are calculated with more accuracy. The horizontal sampling constant h_h can also have variable values if the local curves are further analyzed in order to determine the positions of the states in the diagram in the popular rows that have probability close to 0.

VI. Conclusions

In this paper we proposed a sampling method for size reduction of a state diagram of a stochastic model for peer-assisted VoD streaming in managed networks. Throughout the paper we proved that this method significantly reduces the size of the system, and thus, the number of linear equations that have to be solved for determining the system's performance. This reduction contributes to accelerating the computation time and making the model applicable even for larger networks.

Acknowledgment

The author thanks the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje, under the MultiNets ("Modelling and Analysis of Dynamical Processes in Composite and Multiplex Networks") project for financial support.

References

[1] *Cisco Visual Networking Index: Forecast and Methodology, 2012-2017*, 2013.

[2] Y. Liu, Y. Guo, and C. Liang, "A Survey on Peer-to-Peer Video Streaming Systems," *Peer-to-Peer Networking and Applications,*, vol. 1, no. 1, pp. pp. 18–28, Mar. 2008.

[3] B. Li and H. Yin, "Peer-to-Peer Live Video Streaming on the Internet: Issues, Existing Approaches, and Challenges," *IEEE Communications Magazine,*, vol. 45, no. 6, pp. pp. 94–99, 2007.

[4] Y.-F. Chen, R. Jana, D. Stern, B. Wei, M. Yang, H. Sun, and J. Dyaberi, "Zebroid: Using IPTV Data to Support STB-Assisted VoD Content Delivery," *Multimedia Systems,*, vol. 16, no. 3, pp. 199–214, 2010.

[5] E. Brosh, C. Agastya, and J. Morales, "Serving Niche Video-on-Demand Content in a Managed P2P Environment," *Architecture*, pp. 1–17, 2009.

[6] J. M. Dyaberi, K. Kannan, and V. S. Pai, "Storage Optimization for a Peer-to-Peer Video-on-Demand Network," in *Proceedings of ACM MMSys*, 2010, pp. 59–70.

[7] J. Muñoz Gea, A. Nafaa, J. Malgosa-Sanahuja, and T. Rohmer, "Design and Analysis of a Peer-Assisted VoD Provisioning System for Managed Networks," *Multimedia Tools and Applications,*, pp. 1–36, 2012.

[8] P. Zhu, H. Yoshiuchi, and S. Yoshizawa, "P2P-Based VOD Content Distribution Platform with Guaranteed Video Quality," in *Proceedings of IEEE CCNC*, 2010, pp. 1–5.

[9] A. Korosi, C. Lukovszki, B. Szekely, and A. Csaszar, "High Quality P2P Video-on-Demand with Download Bandwidth Limitation," in *Proceedings of IWQoS*, 2009, pp. 1–9.

[10] K. Kerpez, Y. Luo, and F. J. Effenberger, "Bandwidth Reduction via Localized Peer-to-Peer (P2P) Video," *Digital Multimedia Broadcasting,*, pp. 1–10, 2010.

[11] C. Jayasundara, A. Nirmalathas, E. Wong, and C. A. Chan, "Localized P2P VoD Delivery Scheme with Pre-Fetching for Broadband Access Networks," in *Proceedings of IEEE GLOBECOM*, 2011, pp. 1–5.

[12] S. Gramatikov, F. Jaureguizar, J. Cabrera, and N. García, "Stochastic modelling of peer-assisted VoD streaming in managed networks," *Comput. Netw.*, vol. 57, no. 9, pp. 2058–2074, Jun. 2013.

[13] L. Kleinrock, *Queuing Systems.* Wiley Interscience, 1975, vol. I: Theory.

[14] P. Flajolet and R. Sedgewick, *Analytic Combinatorics.* Cambridge University Press, 2009.

Synchronization and animation of speech in Macedonian language

Zlatko Kuvendjiski *), Boban Joksimovski, Dragan Mihajlov
Faculty of Computer Science and Engineering
St. Cyril and Methodius University
Skopje, Republic of Macedonia
zkuvendz@gmail.com, {boban.joksimoski, dragan.mihajlov}@finki.ukim.mk
*active student at the time of writing

Abstract— **Humans are visual creatures. Roughly 90% of information that is transmitted to the brain is visual, and it is processed 60,000 times faster in the brain than textual information. That is why visualization of speech is such an interesting area for exploration. Seeing someone talk is much more informative than having to read something. Even on presentations, if there is someone (or something) speaking, it holds much more attention for the audience, rather than having the audience just read the same text.**

Lip synchronization is one of the core features in the visualization of speech, along with body language and facial expressions. In computer animation, although it is fairly common, it is one of the more difficult areas to be realistically conveyed, along with accompanying facial expressions.

This work is about generating lip-synced animation from Cyrillic Macedonian text. For that purpose, a topologically correct 3D model of a human face has been created, along with several morph targets for all of the phonemes in the Macedonian language. In addition to that, morph targets for emotions have been created as well, that can render the animation more appealing to the user.

The animation of the model is created by parsing textual information through a python script. It parses the input text and interpunction and creates a corresponding JSON file. The JSON data can be expanded with additional information about the text, like when exactly a given word should be animated, amplitude of the animation, emotions, etc.

It can have multiple applications ranging from creating animations that hearing impaired people can understand, by reading from the lips of the model, to multimedia applications for children. It can be combined with text-to-speech or speech-to-text systems, which, either automatically animate some audio file, or generate a video with sound from a given text.

Keywords—animation; lip-sync; synchronization;

I. Introduction

Facial animation is traditionally considered as an important but tedious task for many applications. Recently the demand for lip-sync animation is increasing, but there are not many fast and easy generation methods. Lip-sync has as steep learning curve.

In the 3D animation field, the quality of productions is continuously increasing. It is a very active market with a high level of competitiveness where modest companies, in terms of budget, must reach a balance between the resources they can apply and the economic investment in a given production. Consequently, the automation of manual design processes, which are normally highly time-consuming, has become a crucial research topic for 3D animation studios. The work we are describing here presents one of these automatic tools, specifically focused on the synchronization of the speech and the lip movement of the characters, a process that is called lip-sync. We have developed a system that reads an input text and generates speech animation based on that. The input can be manipulated to generate different variations of the final animation and to do better synchronization with an audio file.

At first, the phonetics of the Macedonian language will be explained, to show how the animation can be achieved. Then, the process of rigging the model, and generating the blend shapes for the facial expressions and the phonemes will be shown. And in the end, the python script used to generate the final animation.

II. Previous Work

Because this is a very challenging topic, a lot of work has been done so far. While most people cannot read lips (i.e., identify speech from the mouth movement alone [1]), viewers do have a passive notion of correct mouth movement during speech—we know good and bad lip-sync when we see it. In animations where realistic movement is desired, mouth motion and general character movement may both be obtained by rotoscoping [2]. Animation handbooks often have tables illustrating the mouth positions corresponding to a small number of key sounds [3]. The animator must approximately segment the soundtrack into these key sounds. Several viable computer face models have been developed [4,5,6,7]. The goal is to control these face models with a high-level animation script, and have an intelligent front end to the face model automatically translate the script into an appropriate sequence of facial expressions and movements.

III. Macedonian phonology

Phoneme is a voice unit which has meaning in a word. Phonology is the science that studies the phonemes according to their pronunciation and articulation (where and how they are formed and all the speech organs used).

Words can be split into their smallest parts – phonemes. In the Macedonian language there are 31 phonemes.

The phonemes can also be studied according to the function they serve in a given word. For example the word “р а к а” is comprised out of 4 phonemes. If only one is changed, let’s say “р” with “м”, a word will come up, “м а к а”, which has totally different meaning. The phoneme has meaning, if only one phoneme is changed, the meaning of the word will be changed.

A letter is the sign that represents a phoneme. What can be spoken, can also be written with an appropriate sign. In the Macedonian language, each phoneme has its own sign. One letter correlates to one phoneme, which means that there are that many letters as there are phonemes. You write as you speak, and speak as you write. All the phonemes, create the phonetic system, and all the letters, the alphabet.

That makes the Macedonian language very easy to animate. By simply parsing the words in a given text, and splitting the words into letters, a convincing speech animation can be created. That is the whole purpose of this project, to create convincing animation of speech in Macedonian language.

IV. Creation of an Anatomically correct Human model

The model has been created as part of my final graduation exam. It is a photorealistic model of a human head. It has been done using a box modeling method, which produces clean mesh that can be easily animated. This is essential for this project because modeling of the morph targets is crucial for creating convincing animation.

It has been created using Autodesk Maya, using photographed face as a reference (Figure 1), which resulted in very realistic model.

But, realism of the look of the model is not that important, only the realism of the animation. This work can also be done with any high resolution model.

V. Face rigging and control

When the model is created, it's a static 3D mesh, almost like a marble sculpture.

A character rig is essentially a digital skeleton bound to the 3D mesh. Like a real skeleton, a rig is made up of joints and bones, each of which act as a "handle" that animators can use to bend the character into a desired pose.

A character rig can range from simple and elegant to staggeringly complex. A basic setup for simple posing can be built in a few hours, while a fully articulated rig for a feature film might require days or weeks before the character is ready for full feature animation.

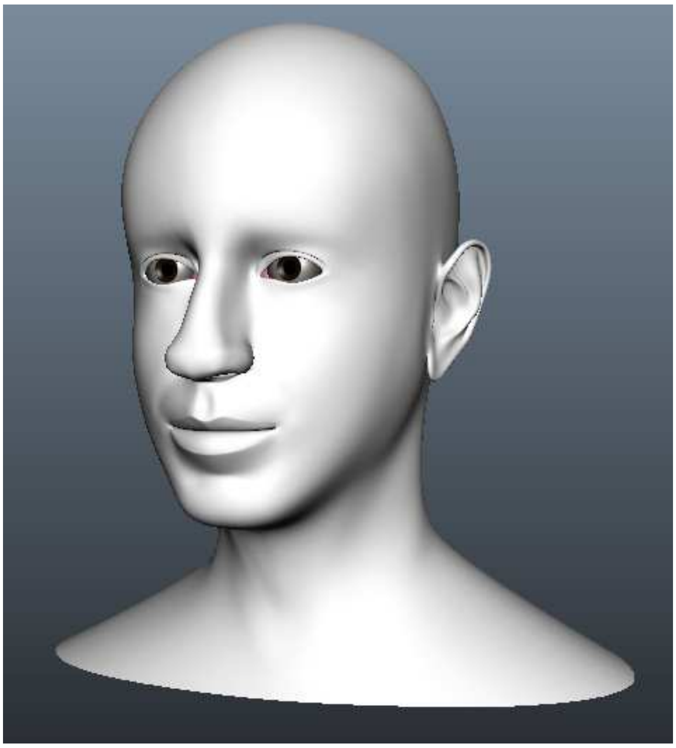


Figure 1. The look of the starting model. All the textures have been removed, because it would interfere with modeling.

The rigging process starts with placing a skeleton. That is perhaps the easiest part of the rigging process. For the most part, joints should be placed exactly where they would be in a real world skeleton, with one or two exceptions.

In order for a rig to work properly, the bones and joints must follow a logical hierarchy. When setting up a character's skeleton, the first joint you place is called the root joint. Every subsequent joint will be connected to the root either directly or indirectly through another joint.

Forward kinematics (FK) is one of two basic ways to calculate the joint movement of a fully rigged character. When using FK rigging, any given joint can only affect parts of the skeleton that fall below it on the joint hierarchy.

For example, rotating a character's shoulder changes the position of the elbow, wrist, and hand. When animating with forward kinematics, the artist typically needs to set the rotation and position of each joint individually—to achieve a desired pose the animator would work through the joint hierarchy sequentially: root → spine → shoulder → elbow → etc. The final position of a terminating joint (like a knuckle) is calculated as a function of the joint angles of every joint above it in the hierarch

As opposed to Forward Kinematics, Inverse Kinematics (IK) is the reverse process, and is often used an efficient solution for rigging a character's arms and legs. With an IK rig, the terminating joint is directly placed by the animator, while the joints above it on the hierarchy are automatically interpolated by the software.

IK is most appropriate when the animation calls for a terminating joint to be placed very precisely, a character climbing a ladder is a good example. Because the character's hands and feet can be placed directly on the ladder's rungs rather than the animator having to adjust their position joint-by-joint, an IK rig would make the animation process far more efficient. One drawback is that because IK animation uses software interpolation, there's often quite a bit of cleanup work that must be done in order to finalize the shot.

This project uses inverse kinematics to create bones and joints for the neck, head, the tongue and the jaw (Figure 2). When the mesh was applied to the bones, a lot of weight painting had to be done on the individual vertices in the mesh to isolate which bone has influence on which part of the mesh. Except for the jaw, the other bones are really not necessary for the facial animation, but can be used to create secondary movements on the model, to achieve even greater effect.

For easier animation, the bones are connected to simple shapes, which drives the bones, and therefore the whole model. For example, the upper bone was connected to the rotation controls of a simple circle. By rotating the circle, the whole head is being rotates in a same way a human being would rotate his head. Also, custom parameters have been added to so we can effectively drive the jaw. With a simple slider, the jaw can be opened or closed.

Using a similar technique, the eyes are rigged to look at simple shapes. By moving those shapes, the eyes move. This also can be used to add realism in the final animation.

A character's facial rig is usually altogether separated from the main motion controls. It is inefficient and incredibly difficult to create a satisfactory facial rig using a traditional joint/bone structure, so morph targets (or blend shapes) are usually seen as a more effective solution.

VI. Creating Expressions by using Blend Shapes

The "morph target" is a deformed version of the model shape. When applied to a human face, for example, the head is first modelled with a neutral expression and a "target deformation" is then created for each other expression. When the face is being animated, the animator can then smoothly morph (or "blend") between the base shape and one or several morph targets. Typical examples of morph targets used in facial animation is a smiling mouth, a closed eye, and a raised eyebrow, but the technique can also be used to morph between whole digital human bodies, for example transforming some character into a creature.

When used for facial animation, these morph targets are often referred to as "key poses". The interpolations between key poses when an animation is being rendered, are typically small and simple transformations of movement, rotation, and scale performed by the 3D software.

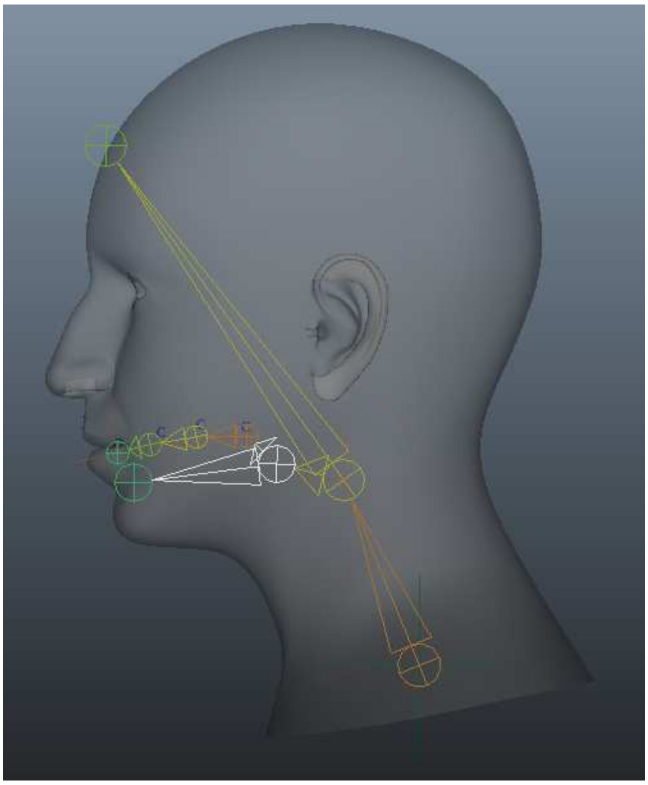


Figure 2. The simple rig for controlling the overall movements of the model.

Blend shapes create the illusion that one shape changes into another in a natural-looking way. It can be used, for example, to animate a character’s mouth moving from a neutral shape into a smile.

This works by using a duplicated version of the object, which is then manually adjusted to another shape. Then, the blend shapes can be used to blend or morph between these shapes, and creates the illusion of an object changing its form.

Before a blend shape can be created, several things have to be taken into considerations about the geometry that should be blended.

The first and possibly the most important of these considerations is the vertex count. In order for a blend shape command to be successful, the base mesh and any target shapes being used must have the same vertex count, otherwise the software will report an error or simply not carry out the operation.

The simplest way to create the blend shapes is by copying the main, or the neutral model. That way, the vertex count stays the same. Each target represents one facial expression. The facial expressions can vary from expressions for different emotions, to an expression for a phoneme, or several phonemes at once.

The emotion expressions of the model vary from a simple smile to frown and sad (one is for the mouth, other for the eyes), wink, angry, astonished and many other. Although they are not used for the speech animation, they are used for adding secondary motion.

The most important shapes created, are the ones for the phonemes. Because some phonemes look alike, the same shapes are used for different phonemes. After a lot of research for similar looking phonemes, and a lot of trial and error, the final shape groups can be divided into:

- А – This is a very specific phoneme, and it is the only one created by only opening the jaw.
- В, Ф – The lower lip goes under the upper one to create this sound.
- Г, Ѓ, К, С, Ј, Ќ – This shape is created by opening the mouth a little bit. The specific sound is created using the middle or the end part of the tongue.
- Д, Н, Л, Р, Т, Ц – This group is similar to the previous. The only difference is that the tongue, touches the teeth with its front part.
- Е – This is also characteristic phoneme, which is partly created by adding a smile shape.
- Ж, Ч, Џ, Ш – This is similar to the "О" phoneme, but with the lips pushed further to the front.
- Ѕ, З – This has a faint smile element in it, with the tongue touching the teeth.
- И – More expressed smile in it.
- Љ, Њ – This is achieved by pressing the middle part of the tongue on the upper gum.
- М, П, Б – The upper and lower lips are pressed on each other and retracted backward a bit.
- О – Very characteristic shape, which is created by rounding the lips.
- У – Similar to the previous one, but the lips are tighter, and more to the front.
- Х – This sound is generated using the tongue and letting air pass over it, therefore the shape is very simple, and is achieved by slightly opening the mouth.

These shapes are then connected to a slider/driver, and by moving the driver, the model morphs into the target shape. The 3D software, automatically interpolates the shapes. So if a slider for a given shape is set to 50%, the model is half and half blended between the target shape and the neutral model. Also, several shapes can be blended at once, producing very interesting results. This is especially great for the final animation where the shapes transform from one to another, producing a realistic animation.

VII. Scripting

Python is a general purpose scripting language used in many different industries. It is a relatively easy to use and easy to learn language. Python is used in internet services, hardware testing, game development, animation production, interface development, database programming, and many other domains. The Python Interpreter is the software package that takes the code and translates it into a form that the computer can understand in order to execute the commands. This translated form is called byte code.

A Python script is simply a collection of commands written in a file with a .py extension. This text file is also called a Python module. This .py file can be written in any text editor like Note Pad, Word Pad, vi, emacs, Scite, etc.

Autodesk Maya supports the use of Python-style scripting or its internal scripting language, MEL. The implementation of Python scripting in Maya provides the same access to

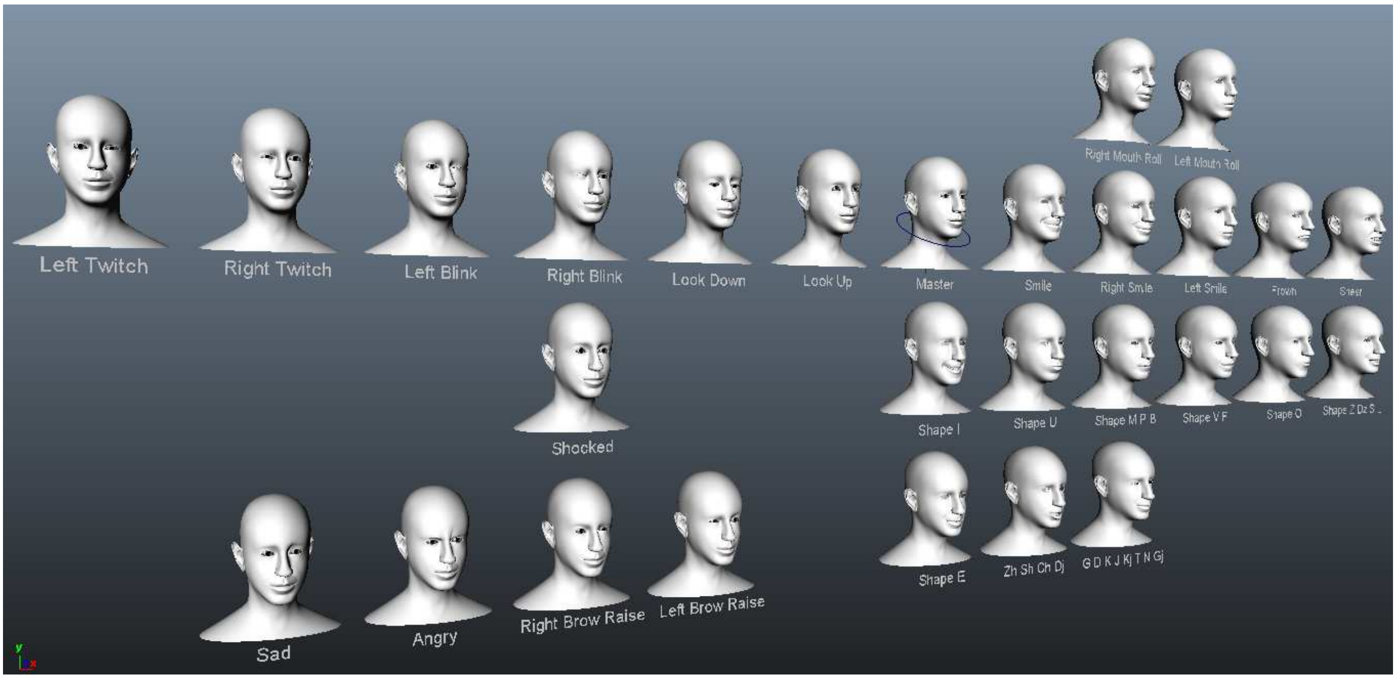


Figure 3. All the blend shapes created. Because the process for defining the correct shapes for the phonemes was iterative, this is not the final state of the shapes created for the phonemes.

native Maya commands as is provided through MEL. That is, all of the built-in Maya commands, such as sphere, ls, and so on, are accessible through Python.

The script created for this project consists of two separate parts. The first part, imports the input text, and creates a new JSON structure. This is required because several optional parameters can be defined for every word separately:

- Amplitude – the blend value for the animation. Default is 0.5, which creates normal looking speech animation. But if it is set to a higher value, it creates more exaggerated animation. Typically used when the character shouts loudly.
- Start time – If the character doesn't speak for a while, the start time of the next word should be offset for that timeframe.
- Speed – If the character speaks some part of a sentence faster, the speed of the animation can be controller with this parameter.

Also, in the JSON file, emotions can be defined with almost the same parameters as the word structures. These can be used to transfer the mood of the speech.

The second part of the script is the actual animation. It reads the JSON file, and creates appropriate keyframes. The result is a natural looking animation of the given speech.

VIII. Further work

This project relies on having a transcribed speech that is used as input for the scripts. That is why this is not completely automated process.

There is a lot of room for upgrading the functionality of this engine. Speech recognition can be integrated to automatically convert an audio file into text and then create an animation.

It can be also done the other way around. If only text is given, speech synthesis engine can be added, and along with the generated animation from our engine, text readers can be created for hearing impaired people.

Unfortunately, engines for that purpose are not yet fully developed for the Macedonian language.

Also, the script can be expanded to make use of the rig structure. For example, randomized head movements can be added, or animate the model to look at the camera, while it pans around it.

All these secondary movements complement each other to create more realistic looking animation.

IX. Conclusion

Speech animation has a wide range of applications, including video game and movie characters, medical visualization, psychological stimuli, online avatars and virtual guides. The most popular technique employed to generate facial animations for movies and video games is blend-shape animation. This allows an animator fine control over a facial model in order to add subtle nuances or correct some unwanted motion.

Performance driven animation is also a popular animation technique, but very often manual work is needed, e.g. to adjust the blend shape weights on a model. High quality animation, such as that seen in movies, is always the result of several techniques and a large proportion of manual effort. Therefore, there is always an interest in the development of new techniques that can increase productivity and reduce development time.

That's why, the idea of being able to automatically generate a facial animation from speech is a highly attractive proposition. Given such a technique, an actor's voice track could be used to automatically animate a facial model, including lip-synching and facial expression. This has advantages over e.g. performance driven animation which additionally involves physically recording an actor's performance using a capture system.

Speech driven animation also has great potential in online video games. In this case, the voice of a person speaking to their friend may be translated onto their virtual avatar. This would far improve current online games where the avatars of two people conversing by voice do not show any sign of motion or interaction, and instead are quite wooden.

References

[1]. E. Walther, Lip-reading, Nelson-Hall, Chicago, 1982.

[2]. T. McGovern, The Use of Live-Action Footage as a Tool for the Animator, SIGGRAPH 87 Tutorial Notes on 3-D Character Animation by Computer, ACM, New York, 1987.

[3]. P. Blair, Animation: Learn How to Draw Animated Cartoons, Foster, Laguna Beach, California, 1949.

[4]. F. Parke, 'Parameterized models for facial animation', IEEE Computer Graphics and Applications, 2, (9), 61-68 (Nov. 1982).

[5]. P. Bergeron and P. Lachapelle, Controlling facial expressions and body movements in the computer generated animated short "Tony de Peltrie", SIGGRAPH 85 Tutorial Notes, ACM, New York, 1985.

[6]. N. Magnenat-Thalmann and D. Thalmann, Synthetic Actors in Computer Generated Three-Dimensional Films, Springer Verlag, Tokyo, 1990.

[7]. K. Waters, 'A muscle model for animating three dimensional facial expression', Computer Graphics, 21, (4), 17-24 (July 1987).

[8]. V. Blanz, and T. Vetter 1999. A morphable model for the synthesis of 3d faces. In Proc. of SIGGRAPH.

[9]. C. Busso, 2007. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. Los Angeles, CA.

[10]. S. Kshirsagar, 2001. Principal components of expressive speech animation. MIRALab CUI, Geneva Univ., Switzerland.

Web-based system for textual retrieval of medical images

Ivan Kitanovski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
ivan.kitanovski@finki.ukim.mk

Ivica Dimitrovski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
ivica.dimitrovski@finki.ukim.mk

Suzana Loskovska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
Suzana.loskovska@finki.ukim.mk

***Abstract*—This paper presents a web-based system for textual retrieval of medical images. The system is built on a scalable architecture using a combination of the latest web technologies and tools. Its main goal is to provide a web interface to the Terrier IR search engine. The interface allows the users to enter the input keywords and also configure the retrieval by a number of configuration options, including query expansion, adding special weights to words, as well as, choosing between different weighting models. The underlying architecture allows the system to be easily modified, since it is highly modular with well defined endpoints between different parts.**

Keywords—textual retrieval; medical retrieval; web-based system

I. Introduction

Health care is a major business in developed countries. It has been reported [1] that 16% of the gross domestic product (GDP) of the United States came from the health care sector in 2004. The numbers were similar for other developed countries: 10.9% in Switzerland, 10.7% in Germany, 9.7% - Canada etc. As technology progresses, the percentage of GDP spent on health care will increase.

Significant part of the health care is related to management and retrieval of medical data [1]. The widespread use of the web has dramatically changed the way people acquire information. Medical professionals are increasingly using web-search engines to acquire more information and to keep up to date with rapid development of the medical knowledge [1].

Hence, search engines play a vital role in today's information society. Search engines specifically tailored for the medical domain are a key tool for physicians as well as researchers. Nevertheless, medical information retrieval has its unique requirements that distinguish itself from traditional information retrieval. In response to these specific needs a number of medical search engines have been launched since 2005. These systems include Healthline [2], Medstory [3], Curbside [4], SearchMedica [5] etc. While these systems have their own merits, they mostly treat the medical search as any other type of search.

The former is also true, when it comes to retrieving medical images. For a given input set of keywords, the search engines treat all input words as equal, not taking into consideration their medical significance. For example, modality is one of the key features provided as input for retrieving images. Having this in mind, one can add additional importance to the keyword when its meaning is related to modality or even more, the user can be allowed to define the his/her own weight distribution which will reflect on the importance on each input keyword. Medical information retrieval system should also perform query expansion by adding multiple related keywords on top of the given input to further widen the result set. Also, allowing the user to switch between different weighting models, might give him/her the ability to fine-tune the final results.

The subject of this paper is the implementation of an interactive web-based system for text-based retrieval of medical images. The web-based system allows the user easy to configure his/her query. The systems allows switching between different weighting models, tuning cost parameters, as well as adding special weights to words which the author deems important. The system integrates query expansion techniques relaying on pseudo relevance feedback. This feature is configurable. The system uses the ImageCLEF 2013 [6] dataset as core database i.e. the provided queries will return results from that set of data.

The rest of the paper is organized as follows: Section 2 presents the related work in the field. The requirements of the systems are presented in 3. Section 4 contains the architecture of the system. The technical details are presented in section 5. Finally, section 6 provides the concluding remarks and future work.

II. Related work

Search engines for the medical domain are an active research area. Already several systems for medical retrieval (visual or textual) are implemented and can be found online. The NovaMedSearch [7] group has created a multi-modal image and case retrieval system over the CLEF 2013 dataset. Their system takes keywords and/or images as input and returns a sorted set of 10 images or cases. The system provides an assisted query expansion functionality which basically provides auto-completion on the input terms and automatic query expansion with semantically related terms. Although, user-friendly, the system lacks the possibility to configure in more detail the underlying retrieval techniques.

PubMed [8] is a specialized search engine, designed to search for publication information (also referred to as citation) over the MEDLINE [9] database. This is a very powerful tool for medical professionals; based purely on the textual input of the user. The user can specify which fields of the article can/must contain a keyword and combine them within boolean expression. Although, useful in medical practice it does not have the ability to retrieve only medical images.

The MedSearch [1] search engine developed by IBM T.J. Watson Research Center is a specialized medical Web search engine, which uses several key techniques to improve usability and the quality of search results. It allows queries of extended length and reformats long queries onto shorter ones by extracting a subset of important and representative words. It allows suggestions of medical phrases to be added in the input, since the system is meant for non-professional users. Nevertheless, the system as a black box lacks the possibility to manually redefine the weighting distribution of the input set of keywords.

The MedGift group [10] has designed two search engine interfaces for demonstration purposes: text based case retrieval search engine] and a visual images search engine. These systems perform well in their respected domains, but for instance there is no way of retrieving images by means of a textual query.

To overcome the problems of other systems, we propose a system which adds value by allowing users to search for medical images in particular by means of a textual query.

III. System requirements

Developing a medical retrieval system for images has its own unique challenges, since we can use domain knowledge in the way the images are represented, indexed and retrieved. We must have in mind that there are several key features that the medical retrieval system must fulfil:

1. The system requires access to an image repository and optionally a meta data repository
2. The system requires a method of making representations the images in the repository and indexing it
3. It must provide the user with an interface to enter the input keywords for the retrieval
4. It must provide the user the ability to configure the retrieval by different input parameters
5. The system must have a mechanism to calculate a score of relevancy of the images according to the given input
6. The system must provide the user with a list of medical images sorted by relevancy
7. The system should be highly responsive to user actions
8. It must be scalable and able to work with large amounts of data

Having these requirements in mind, we defined a architecture using the latest web technologies and state of the art tools.

IV. Architecture

The system uses a standard three tier architecture. Each tier is independent from the others and can be easily replaced with another. The system architecture is depicted on Figure 1.

The front-end server is a web application which is in charge of handling user requests and sends those requests to the application server. Basically, it is a proxy and load balancer. This part of the system performs initial parameter validation of the parameters sent to the application server. After the parameters are sent and the job is completed, the results are presented to the user.

The application server is the part of the system which executes a retrieval based on a set of given input parameters. It provides its services through a RESTfull [11] architecture, which in turn allows to be used by client through HTTP requests. The application server acts as a wrapper to the underlying search engine. It validates and converts the parameters into the format the search engine recognizes and passes them. Once, the job is done by the search engine a sorted list of items is returned via the API.

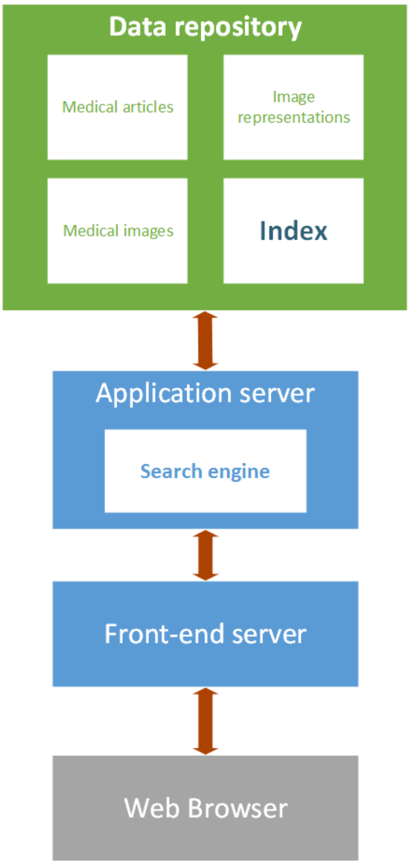


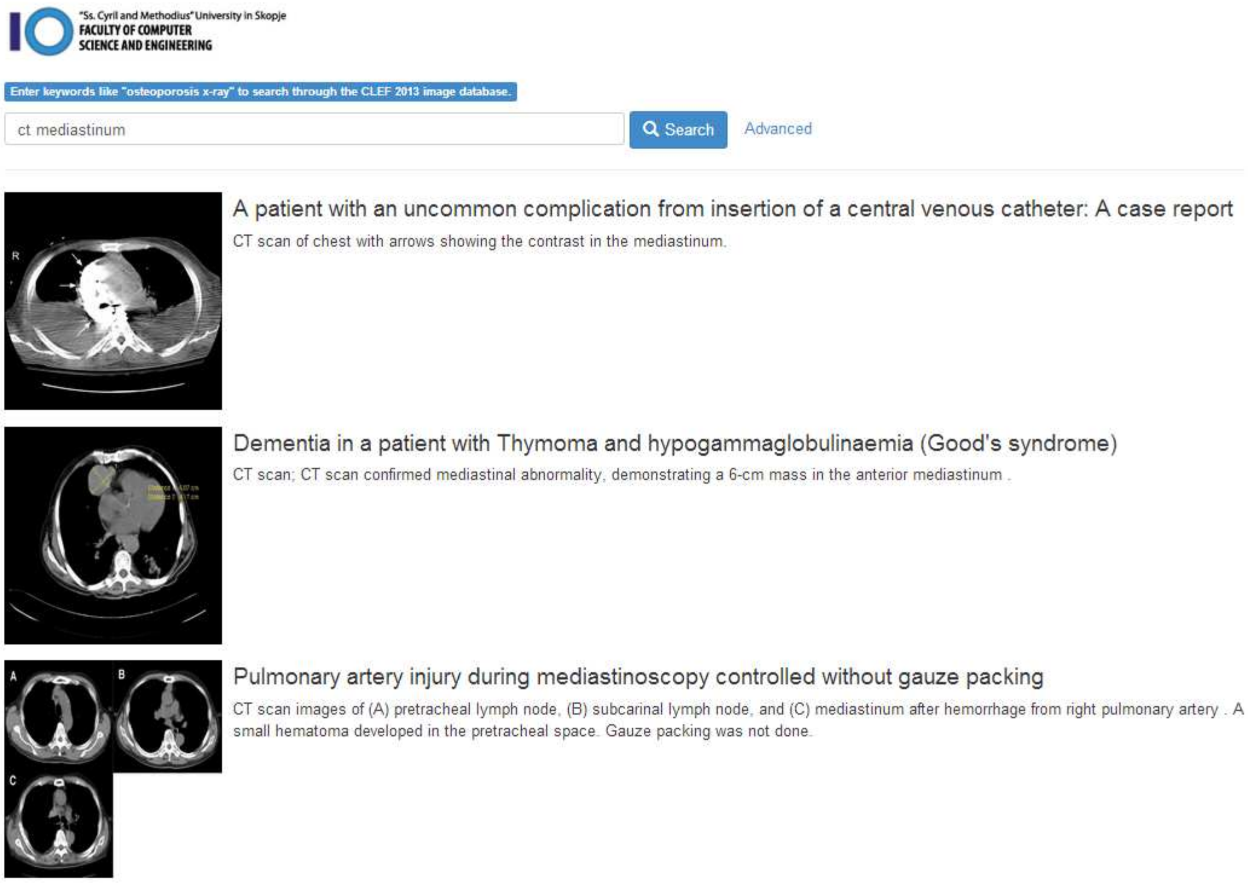
Fig. 2. System architecture

The data repository contains the index created by the search engine, the medical images, textual representations for the images provided from the medical articles where images are referenced. The data repository is kept completely on the file system. The application server accesses it directly through the system.

V. Technical details

The system is composed of a variety of technologies, each used in a different part of the system.

The front-end part of the system is developed in the



"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Enter keywords like "osteoporosis x-ray" to search through the CLEF 2013 image database.

ct mediastinum

A patient with an uncommon complication from insertion of a central venous catheter: A case report

CT scan of chest with arrows showing the contrast in the mediastinum.

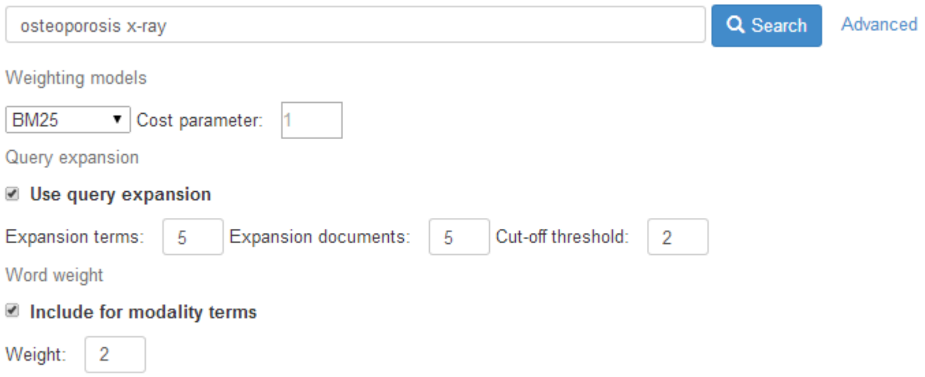
Dementia in a patient with Thymoma and hypogammaglobulinaemia (Good's syndrome)

CT scan; CT scan confirmed mediastinal abnormality, demonstrating a 6-cm mass in the anterior mediastinum .

Pulmonary artery injury during mediastinoscopy controlled without gauze packing

CT scan images of (A) pretracheal lymph node, (B) subcarinal lymph node, and (C) mediastinum after hemorrhage from right pulmonary artery . A small hematoma developed in the pretracheal space. Gauze packing was not done.

Fig. 2. System UI



osteoporosis x-ray

Weighting models

BM25 ▾ Cost parameter: 1

Query expansion

☑ **Use query expansion**

Expansion terms: 5 Expansion documents: 5 Cut-off threshold: 2

Word weight

☑ **Include for modality terms**

Weight: 2

Fig. 3. System configuration

Python's Flask [12] micro-framework for web development. For the overall web page layout and interactivity the system uses Twitter's Boostrap [13] and jQuery [14] framework. The UI of the system can be seen on Figure 2. The configuration options are displayed on Figure 3.

The back-end (application server) part of the system is a Java web application which provides its RESTfull services using the Jersey framework [15]. It is a wrapper for the Terrier IR search engine [16]. When the Java side sends parameters to Terrier for retrieval it receives a sorted list of image representations as a result. The Java side then maps the image representations to actual images and returns that as a result to the front-end.

The Terrier search engine creates and keeps its index in a custom file structure and requires pointers to the files that need to be indexed. Since, the data is kept entirely on the file system, this does not pose a problem. In our case the files that should be indexed are the textual representations of the images. The images are also kept on the file system.

VI. Conclusion

We have built a web-based system which allows retrieval of medical images based on a textual input. The system serves as an interface to the underlying Terrier IR search engine and its main goal is to enable a user to configure the retrieval by multiple parameters, rather then just by the input keywords. The system currently uses only one repository, but can be easily extended with more data stores. One of our future goals is to add a module for retrieval of medical articles (cases). This will add even more value to the system by widening its application domain.

Acknowledgement

This work is partially supported by the Faculty of Computer Science and Engineering, Skopje, Macedonia as a part of the project “Scalable Photo Annotation”.

References

[1] G. Luo, et al. “MedSearch: A Specialized Search Engine for Medical Information Retrieval”, Proceedings of the 17th ACM conference on Information and knowledge management, pp. 143-152, 2008.

[2] [http://www.healthline.com/] last visited: 29.03.2014

[3] Anderson, Patricia F. "Medstory." Journal of the Medical Library Association 95, no. 2, pp. 221, 2007

[4] [http://www.curbside.md/] last visited: 29.03.2014

[5] [http://www.searchmedica.co.uk/] last visited: 29.03.2014

[6] A. G. Seco de Herrera, et al. "Overview of the ImageCLEF 2013 medical image retrieval and classification tasks," Working Notes of CLEF, 2013.

[7] A. Mourao, et al. "NovaSearch on medical ImageCLEF 2013" Working Notes of CLEF, 2013

[8] [http://www.ncbi.nlm.nih.gov/pubmed] last visited: 29.03.2014

[9] D. A Lindberg,"Internet access to the National Library of Medicine." Effective clinical practice: ECP, 2000.

[10] A. G. Seco de Herrera, et al. “The medGIFT group in ImageCLEFmed 2013”, Working Notes of CLEF, 2013.

[11] Alex Rodriguez,. "Restful web services: The basics." Online article in IBM DeveloperWorks Technical Library 36, 2008.

[12] [http://flask.pocoo.org/] last visited: 29.03.2014

[13] [http://getbootstrap.com/] last visited: 29.03.2014

[14] [http://jquery.com/] last visited: 29.03.2014

[15] [https://jersey.java.net/] last visited: 29.03.2014

[16] I. Ounis, G. Amati, and V. Plachouras, “Terrier information retrieval platform,” Advances in Information Retrieval, pp. 517-519 2005.

Session 2

Interdisciplinary Research

Bioinformatics Cloud Applications

Monika Simjanoska, Marjan Gusev and Ana Madevska Bogdanova
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Email: m.simjanoska@gmail.com, {marjan.gushev, ana.madevska.bogdanova}@finki.ukim.mk

***Abstract*—Cloud computing offers an ideal storage and computation opportunities to the scientists who want to migrate their applications in the cloud. In this paper we present various of cloud-based frameworks and tools that are developed to solve high-demanding bioinformatics problems in the field of comparative genomics, DNA sequencing, microarray data analysis, drug analysis, etc. Furthermore, we explore the completely new approaches for improving the performance when running experiments, parallel and distributed solutions, as well as case studies on cost-effective and flexible cloud computing. Evaluating the published materials is of great advantage for our future research in the field of microarray gene expression analysis accompanied by the development of platform for colorectal cancer research in the cloud.**

Keywords*—*Bioinformatics, Cloud Computing, Scientific Tools

I. Introduction

Cloud computing is a very convenient opportunity to get advantage of highly memory-demanding and computation-intensive bioinformatics problems. Stein in his research on cloud computing in genome informatics [1] presents how the DNA sequencing is getting cheaper compared to data storage or computation, and emphasizes that the time may have come for genome informatics to migrate to the cloud. In Figure 1 he presents the old genome informatics ecosystem where the sequencing laboratories transmit sequencing data across the internet to one of several sequencing archives. Then the archived information can be accessed either directly by casual users or indirectly via a website run by one of the value-added genome integrators. Power users typically download large datasets from the archives onto their local compute clusters for further analysis. Under this model, the sequencing archives, value-added integrators and power users maintain their own compute and storage clusters and keep local copies of the sequencing datasets.

This system worked perfectly until the doubling time for DNA sequencing was slower than the growth of compute and storage capacity. However, in 2005 the new DNA sequencing technologies caused the costs of sequencing to decrease several times compared to the costs for storage. Then the traditional systems confronted a possible tsunami of genome data that was impossible to be handled by the current storage systems. Therefrom comes the new genome informatics ecosystem based on cloud computing, presented in Figure 2.

In this ecosystem, all the compute and storage resources are located in the cloud which is maintained by a cloud service provider. The sequence archives and value-added integrators maintain servers and storage systems within the cloud. Casual users access the data via the websites of the archives and integrators. The power users now have the option of creating virtual on-demand compute clusters within the cloud, which have direct access to the sequencing datasets.

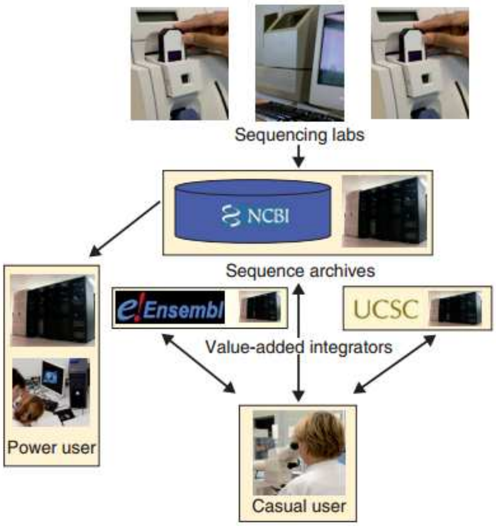


Fig. 1. The old genome informatics ecosystem [1]

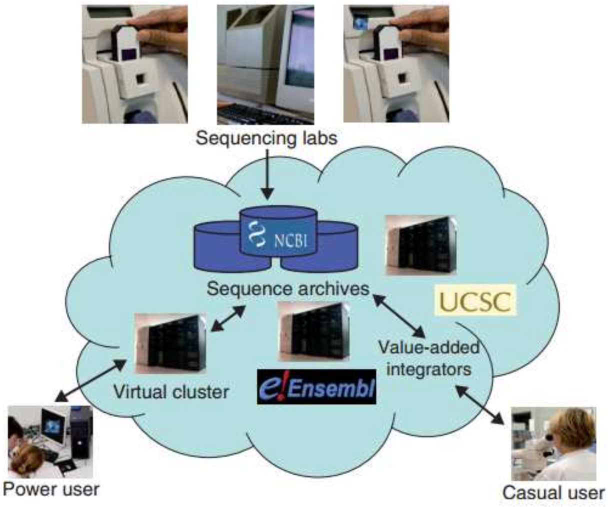


Fig. 2. The new genome informatics ecosystem [1]

In [2] Rosenthal et al. present their cloud computing analysis of whether and when is appropriate for the biomedical laboratories to migrate to the cloud. In Figure 3 they describe a traditional, generic, laboratory infrastructure (left) and a cloud laboratory infrastructure (right).

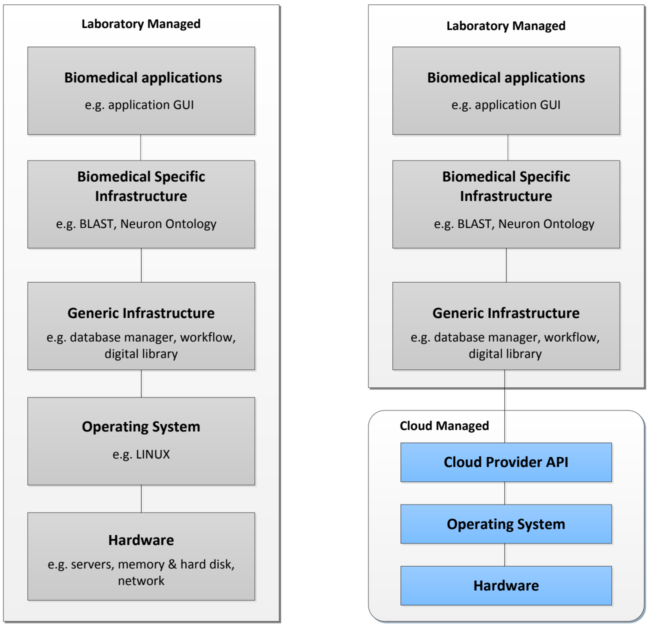

Laboratory Managed
Biomedical applications
e.g. application GUI
Biomedical Specific Infrastructure
e.g. BLAST, Neuron Ontology
Generic Infrastructure
e.g. database manager, workflow, digital library
Operating System
e.g. LINUX
Hardware
e.g. servers, memory & hard disk, network
Laboratory Managed
Biomedical applications
e.g. application GUI
Biomedical Specific Infrastructure
e.g. BLAST, Neuron Ontology
Generic Infrastructure
e.g. database manager, workflow, digital library
Cloud Managed
Cloud Provider API
Operating System
Hardware

Fig. 3. Generic and cloud computing infrastructure

As illustrated on the left in Figure 3, the basic resources as computation, storage and network bandwidth are managed by an operating system. On top of the operating system exists a complex generic infrastructure as database management, digital library, etc. The advantage of the generic infrastructure is used by the biomedical specific infrastructure (such as BLAST-Basic Local Alignment Search Tool). Eventually, on top of all layers come the deployed biomedical applications.

The right side of Figure 3 presents exactly the same infrastructure, but in a case when a cloud is responsible for providing the lower-level capabilities, which means that it is a base for building biomedical applications. The image analysis, data mining, protein folding and gene sequencing are examples of important tools for biomedical researchers and also examples of high variable service demands. Therefore, the authors in [2] sum up the factors that make one biomedical project an attractive candidate for cloud computing which are:

- high costs for computing, administration, space, and electric power;
- secure and rapid data sharing;
- the project requires highly variable amounts of processing and storage resources;
- the system requires off-site backups for data and for processing;
- the applications have easily parallelized code and
- necessity of long-term repositories to outlive the laboratory that now hosts the data.

One the other hand, there are likely obstacles to occur due to the hardware demands of the current tool. For example, some HPC applications as protein folding and high-end image processing, in order to work, require detailed knowledge of the underlying hardware which is not always revealed by the vendor. Those applications are expected to perform poorly in the cloud. Also, there are often technical issues as software portability. Therefore, the customers should seek a cloud service provider that offer most common Unix, Linux or Windows servers. Cloud unfamiliarity and immaturity can also be a problem in maintaining a security of the sensitive data in the virtual machines. The overall conclusion of their research show that transitioning to a cloud will change the ways in which biomedical systems are built, managed and funded. Considering the costs they point out that even a rough analysis will point toward clouds.

The rest of the paper is organized as follows. In Section II we present some of the most common tools and platforms developed for solving high throughput biomedical problems. Discussion about whether the cloud is performing as good as expected is presented in Section III. Finally, in Section IV we derive a conclusion from this survey and present our plans for future work.

II. Bioinformatics Applications and Frameworks in the Cloud

In this section we present various tools and frameworks developed for bioinformatics analysis in the cloud.

A. *FX*

Hong et al. [3] present their tool for RNA-Seq analysis, FX, which runs in local Hadoop systems as well as the Amazon cloud system, for the estimation of gene expression levels and genomic variant calling. In the mapping of short RNA-Seq reads, FX uses a transcriptome-based reference primarily, generated from 160 000 mRNA sequences from RefSeq, UCSC and Ensembl databases. This approach reduces the misalignment of reads originating from splicing junctions. Unmapped reads not aligned on known transcripts are then mapped on the human genome reference. FX allows analysis of RNA-Seq data on cloud computing infrastructures, supporting access through a user-friendly web interface.

B. *YunBe*

Another tool that uses Amazon's cloud infrastructure is YunBe, developed for gene set analysis in the cloud [4]. The tool uses a specially designed algorithm for biomarker identification in the cloud. It is written in Java and uses MapReduce framework to parallelize the analysis. To test YunBe they analyzed published and simulated gene expression datasets. The liver dataset contained 466 samples from 31,842 transcripts. The YunBe's execution speeds has been compared to a program version running on a computing cluster, which consisted of dual socket quad-core Intel E5430 Harpertown CPUs. In this analysis, 1, 2, 4 and 8 Amazon's EC2 m1.large instances have been compared with 2, 4, 8 and 16 cluster cores running at BGI. Comparisons have also been made with a desktop program running on a duo-core Intel E7500 Wolfdale CPU. The results in Figure 4 showed that in comparison to a desktop implementation, YunBe significantly improves execution times.

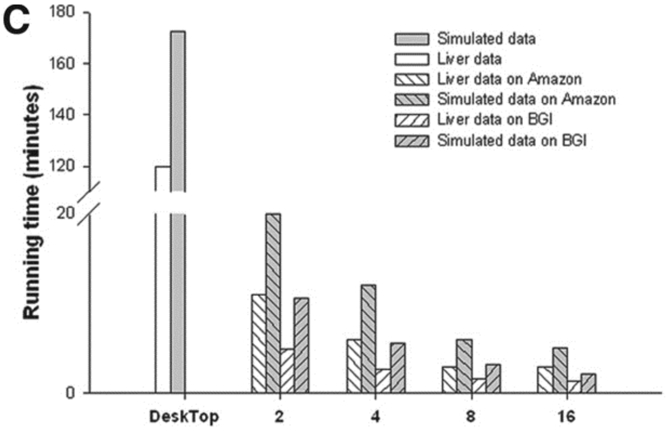


Fig. 4. YunBe improvements [4]

C. CloVR

CloVR (Cloud Virtual Resource) [5] relies on two enabling technologies, virtual machines (VMs) and compute clouds, to provide improved access to bioinformatics workflows and distributed computing resources. CloVR provides a single VM containing pre-configured and automated pipelines, suitable for easy installation on the desktop and with cloud support for increased analysis throughput. In building the CloVR VM, they have addressed the following technical challenges in using cloud computing platforms: elasticity and ease-of-use, limited network bandwidth and portability. The architecture of CloVR addresses these challenges by simplifying use of cloud computing platforms by automatically provisioning resources during pipeline execution, using local disk for storage and avoiding reliance on network file systems, and providing a portable machine image that executes on both a personal computer and multiple cloud computing platforms. The authors in their paper evaluate the features of the CloVR architecture as portability across different local operating systems and remote cloud computing platforms, support for elastic provisioning of local and cloud resources, scalability of the architecture and use of local data storage on the cloud. Their test cases showed that CloVR VM and associated architecture lowers the barrier of entry for utilizing complex analysis protocols on both local single and multi-core computers and cloud systems for high throughput data processing.

D. Cloud BioLinux

Cloud BioLinux [6] provides a platform for developing bioinformatics infrastructures on the cloud. It is a publicly accessible Virtual Machine that enables scientists to quickly provision on-demand infrastructures for high-performance bioinformatics computing using cloud platforms. Users have instant access to a range of pre-configured command line and graphical software applications, including a full-featured desktop interface, documentation and over 135 bioinformatics packages for applications including sequence alignment, clustering, assembly, display, editing, and phylogeny. Besides the Amazon EC2 cloud, they have started instances of Cloud BioLinux on a private Eucalyptus cloud and demonstrated access to the bioinformatics tools interface through a remote connection to EC2 instances from a local desktop computer. The authors compared their platform to another projects like CloVR, Bioconductor, Qiime and GMOD that demonstrate the usefulness of pre-configured VMs with specific types of analysis for biologists and stated that by establishing Cloud BioLinux as a community framework they hope to ease the production of these resources for a broad audience of bioinformatics developers from various backgrounds and with different research goals. Even though they have successfully completed data analysis, measurements of performance and costs for different cluster setups is not further discussed.

E. BioVLAB

BioVLAB [7] is a virtual analysis system for microarray gene expression data in computing clouds with flexible and configurable GUI workflow engine so that biologists are able to analyze the data in many angles without worrying about computational and bioinformatics issues. The authors claim that the contribution of their system is three-fold:

- providing a suite of microarray analysis applications which can utilize remote high-performance computing resources such as computing clouds or public Web services,
- providing an easy-to-use and reconfigurable workflow system in which a workflow composition requires no system knowledge of working environment and users can repeatedly execute the same workflow with different parameter settings, and
- building a Web portal where an administrator can manage inventories of applications that a user can use for a workflow composition and also users can manage their data.

In their paper, they present graphical experiment summary of the performed gene expression analyses.

F. CloudMan

The authors in [8] present a cloud resource management system, CloudMan, that makes it possible for individual researchers to compose and control an arbitrarily sized compute cluster on Amazons EC2 cloud infrastructure without any informatics requirements. CloudMan is built on top of a Bio-Linux machine image available from Cloud BioLinux (previously discussed) and thus makes all of the tools packaged by NERC Bio-Linux immediately available. In addition to the tools available through Bio-Linux, a set of NGS tools available through Galaxy are also available for use, including: Bowtie, BWA, and SAMtools. If a user desires additional tools, they have provided a mechanism for streamlining the tool installation process. The provided solution makes it possible, using only a web browser, to create a completely configured compute cluster ready to perform analysis in less than five minutes. Moreover, they provide an automated method for building custom deployments of cloud resources.

G. Eoulsan

Eoulsan [9] is a scalable framework based on the Hadoop implementation of the MapReduce algorithm dedicated to high-throughput sequencing data analysis. Eoulsan has been developed in order to automate the analysis of a large number of samples at once, simplify the configuration of the cloud computing infrastructure and work with various already available analysis solutions. Eoulsan can be run under three modes: standalone, local cluster or cloud computing on Amazon Elastic MapReduce (EMR). The software has been tested on AWS for a total of 188 millions reads using different read mappers embedded in the workflow. They estimated the time needed to perform the calculation process and the cost charged by AWS on three different EC2 virtual machine types: m1.large, m1.xlarge and c1.xlarge. The fastest result is obtained from the c1.xlarge instance whatever the mapper used. In terms of costs, the computation is always more expensive using m1.xlarge instances with m1.large remaining the most economical choice. Testing the speed up of the calculation process by using a large number of computers at once, showed that the number of instances can be increased in order to speed up the data analysis process without the risk to fall in a suboptimal configuration. Finally, they assessed the impact of an increase in raw data on the computation time by running Eoulsan with 16 and 32 samples of 23.5 millions of reads each, respectively, 376 and 752 millions of total reads. The results showed that the relationship between running time and number of samples is also linear which demonstrates that Eoulsan is able to handle the increase in raw data coming from future evolutions of Illumina sequencing devices.

H. GenomeSpace

Eventually, we briefly present GenomeSpace [10] which is a cloud-based interoperability framework to support integrative genomics analysis through an easy-to-use Web interface. GenomeSpace provides access to a diverse range of bioinformatics tools, and bridges the gaps between the tools, making it easy to leverage the available analyses and visualizations in each of them. The tools retain their native look and feel, with GenomeSpace providing frictionless conduits between them through a lightweight interoperability layer. GenomeSpace does not perform any analyses itself; these are done within the member tools wherever they live desktop, Web service, cloud, in-house server, etc. Rather, GenomeSpace provides tool selection and launch capabilities, and acts as a data highway automatically reformatting data as required when results move from the output of one tool to input for the next. GenomeSpace hosts variety of tools and data sources that provide a wide spectrum of genomic analysis and bioinformatics capabilities, as: Cytoscape, Galaxy, GenePattern, Genomica, IGV, UCSC Table Browser, InSilico DB, Cistrome, geWorkbench, ArrayExpress, Gitools, ISAcreator, MSigDB Online Tools, Synapse, etc.

After we presented a few of the many bioinformatics tools and frameworks in the cloud, in the next section we will discuss the scientists' experiences of using the cloud as an appropriate technology for developing the applications.

III. Discussion of Cloud's Leverage for Bioinformatics Problems

In this section we will discuss various experiences of scientists that used cloud technologies for solving biomedical problems.

Even though cloud computing is expected to be good opportunity for large scale analysis, Dudley et al. [11] present a case study where there are limitations for applications in the domain of high-throughput sequence data analysis. In their research they evaluate the use of cloud computing technologies for a translational bioinformatics analysis of a large cancer genomics data set composed of matched replicate SNP genotype and gene expression microarray assay samples for 311 cancer cell lines, comprising 929 gene expression microarray samples and 622 SNP genotype array samples. They suggest that the data analysis illustrated by this case study is characteristic of computational challenges that might be faced by modern clinical researchers who have access to inexpensive high-throughput genomic assay technologies for profiling their patient populations. The problem they work on was motivated by a need to discover cancer-associated eQTLs through integration of two high-dimensional genomic data types (gene expression and genotype), requiring more than 13 billion distinct statistical computations. Considering the SNP platform used to generate the data measured 500,568 SNPs, and that the gene expression microarray platform measured gene expression levels across 54,675 probes, requiring statistical evaluation of more than 13 109 comparisons, they estimated that it would take a single, modern server-class CPU more than 5,000 days to complete the analysis. Hereupon, they demonstrate the computational and economical characteristics of conducting this analysis using a cloud-based service, and contrast these characteristics with the computational and economic characteristics of performing the same analysis on a local institutional cluster. For a cloud computing analysis they used Amazon Web Services (AWS) elastic compute cloud (EC2) and a total of 100 EC2 instances. For the local cluster analysis they used a 240 core High Performance Compute Cluster based on the Hewlett Packard C-class BladeSystem attached to 15 TB storage area network. After executing the analysis, they derived the following conclusions:

- they did not find significant difference in the running time between the cloud and the local cluster;
- total cost for running the analysis on the cloud-based system was approximately three times more expensive compared to the local cluster; however, the cloud solution is still more attractive since there are no start-up costs associated with the cloud-based analysis, compared to the substantial costs for building a local cluster;
- the cloud-based system offers many technical features and capabilities that are not matched by the local cluster, e.g. the cloud's elasticity which allows it to scale the number of server instances based on the need;
- the cloud allows whole systems to be archived to persistent storage for subsequent reuse and 'elastic' disk storage that can be dynamically scaled based on real-time storage needs;
- another feature that may have increased the total execution time of the analysis is launching the new instance during periods of reduced cloud activity; however, it might also reduce the cost of the cloud-based analysis by half depending on market conditions.

They also state some considerations using the cloud:

- cloud computing allows free configuration of virtual machine instances, thereby sharing the burden of security with the user;
- cloud computing requires the transfer of data, which introduces delays and can lead to substantial additional costs given the size of many data sets used in translational bioinformatics;
- an additional limitation they faced was a 1TB limit on the size of the virtual disks;
- the most significant impediment facing biomedical researchers wishing to adopt cloud computing involves the software environment for designing the computing environment and running the experiments, that is cloud-based tools should be specifically oriented to address the particular modes of inquiry of clinician scientists towards enabling unified biological and clinical hypothesis evaluation, instead of offered as a collection of bioinformatics tools (toolbox).

By demonstrating their research, the authors point towards the creation of open-source software tools that take advantage of the cloud computing's features to allow for uploading, storage, integration and querying across large repositories of public and private molecular and clinical data.

Dennis et al. in [12] represent their experience in one of the first successful deployments of a standard comparative genomics tool, the reciprocal smallest distance algorithm (RSD), to Amazon's Elastic Compute Cloud (EC2) via the web service Elastic MapReduce (EMR). They ran more than 300,000 RSD-cloud processes within the EC2. The jobs were farmed simultaneously to 100 high capacity compute nodes using the Amazon Web Service Elastic Map Reduce and included a wide mix of large and small genomes. The total computation time took just under 70 hours and cost a total of $6,302. As a conclusion from their testing they state that the effort to transform existing comparative genomics algorithms from local compute infrastructures is not trivial. However, the speed and flexibility of cloud computing environments provides a substantial boost with manageable cost. The procedure designed to transform the RSD algorithm into a cloud-ready application is readily adaptable to similar comparative genomics problems.

Kudtarkar et al. [13] optimize the computation of a large-scale comparative genomics resource, Roundup, using cloud computing, describe the proper operating principles required to achieve computational efficiency on the cloud, and detail important procedures for improving cost-effectiveness to ensure maximal computation at minimal costs. Utilizing the comparative genomics tool, Roundup, they computed orthologous (diverged genes after a speciation event) relationships for 245,323 genome-to-genome comparisons on Amazons EC2, a computation that required just over 200 hours and cost $8,000, at least 40% less than expected under a strategy in which genome comparisons were submitted to the cloud randomly with respect to runtime. For managing the ortholog processes, they designed a strategy to deploy Elastic MapReduce web service and maximize the use of the cloud while simultaneously minimizing costs. Specifically, they designed a model to predict job runtime and costs for an array of cloud cluster sizes, and showed how this model can be used to identify the optimal cluster size as well as the best strategy for ordering jobs prior to submission to the cloud. Most importantly, they showed how their model can be used to achieve at least a 40% reduction in overall cloud computing costs. The cost-reduction model is readily adaptable for other comparative genomics tools and potentially of significant benefit to labs seeking to take advantage of the cloud as an alternative to local computing infrastructure. In summary, their case study indicates that the cloud is a viable solution for boosting large-scale projects like Roundup, and provides a best-practice model for cloud computing that can be adapted to similar comparative genomics algorithms.

Another case study presented by Wilkening et al. [14] investigate the performance of BLAST on real metagenomics data in a Amazon's EC2 cloud setting, in order to determine the viability of this approach. BLAST is considered to be one of the leading applications in bioinformatics and computational biology and is assumed to consume the vast majority of resources in that area. Their feasibility study on the use of cloud resources in the MG-RAST (meta Genome Rapid Annotation using Subsystem Technology) workflow, indicated the following issues:

- costs are slightly higher to perform computations in the cloud, when compared with local costs;
- the pricing of on-demand resources blunts much of the benefit of EC2s elasticity and
- some security concerns remain to be completely addressed.

However, in several areas a more restricted use of cloud computing could be useful:

- By using cloud resources as a scale-out pool for high-priority jobs, time to solution could be greatly reduced for important jobs. These improvements would come at a substantial cost, however, as the use of on-demand resources incur a large cost penalty.
- If a looser federation mechanism were used for cloud computational instances, users could associate their cloud instances with MG-RAST, providing direct support for their computations. Users would benefit from increased priority for their jobs.

Gunarathne et al. [15] have demonstrated that clouds offer attractive computing paradigms for three loosely coupled scientific computation applications: Cap3 sequence assembly (to assemble a large collection of genome fragments), GTM and MDS interpolation (to perform dimension reduction on 166-dimensional dataset containing 26 million data points obtained from the PubChem project database). They used Amazon Web Services and Microsoft Windows Azure cloud computing platforms and also Apache Hadoop Map Reduce and Microsoft DryadLINQ as the distributed parallel computing frameworks. The results showed that the higher level MapReduce paradigm offered a simpler programming model. Also by using two different kinds of applications they showed that selecting an instance type which suits the given application can give significant time and monetary advantages. The cost effectiveness of cloud data centers combined with the comparable performance reported in their research suggests that loosely coupled science applications will increasingly be implemented on clouds and that using MapReduce frameworks will offer convenient user interfaces with little overhead.

Taylor [16] in his research gives an overview of the current usage within the bioinformatics community of Hadoop and of associated open source software projects. The main focus is on next-generation sequencing. There are many bioinformatics applications, apart from gene-sequencing analysis, that use Hadoop and HBase. Hadoop and its associated open source projects have a diverse and growing community in bioinformatics of both users and developers, as can be seen from the large number of projects. As a conclusion from his work, the author states that for much bioinformatics work not only is the scalability permitted by Hadoop and HBase important, but also of consequence is the ease of integrating and analyzing various large, disparate data sources into one data warehouse under Hadoop, in relatively few HBase tables.

CloudBLAST [17] is an approach that combines MapReduce and virtualization on distributed resources for bioinformatics applications. An implementation of this approach integrates Hadoop, Virtual Workspaces, and ViNe as the MapReduce, virtual machine and virtual network technologies, respectively, to deploy the commonly used bioinformatics tool NCBI BLAST on a WAN-based test bed consisting of clusters at two distinct locations. This WAN-based implementation was evaluated against both non-virtualized and LAN-based implementations in order to assess the overheads of machine and network virtualization, which were shown to be insignificant. To compare the proposed approach against an MPI-based solution, CloudBLAST performance was experimentally contrasted against the publicly available mpiBLAST on the same WAN-based test bed. Both versions demonstrated performance gains as the number of available processors increased, with CloudBLAST delivering speedups of 57 against 52.4 of MPI version, when 64 processors on 2 sites were used. The results encourage the use of the proposed approach for the execution of large-scale bioinformatics applications on emerging distributed environments that provide access to computing resources as a service.

IV. Conclusion and Future Work

In this paper we give a review of the most widely used bioinformatics applications and platforms which aim to solve high-demanding bioinformatics problems in the field of comparative genomics, DNA sequencing, microarray data analysis, drug analysis, etc. All the presented tools are built to take advantage of the cloud computing. Hereupon, we continued our research and presented various case studies with the purpose to answer the question whether the cloud is an appropriate technology for achieving both high performance and cost-effective analysis.

Evaluating the published applications is of great advantage for our future research in the field of microarray gene expression analysis, since our aim is to develop a platform in the cloud that will host tools for colorectal cancer research.

References

[1] L. D. Stein *et al.*, "The case for cloud computing in genome informatics," *Genome Biol*, vol. 11, no. 5, p. 207, 2010.

[2] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester, and P. Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 342–353, 2010.

[3] D. Hong, A. Rhie, S.-S. Park, J. Lee, Y. S. Ju, S. Kim, S.-B. Yu, T. Bleazard, H.-S. Park, H. Rhee *et al.*, "Fx: an rna-seq analysis tool on the cloud," *Bioinformatics*, vol. 28, no. 5, pp. 721–723, 2012.

[4] L. Zhang, S. Gu, Y. Liu, B. Wang, and F. Azuaje, "Gene set analysis in the cloud," *Bioinformatics*, vol. 28, no. 2, pp. 294–295, 2012.

[5] S. V. Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White, and W. F. Fricke, "Clovr: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC bioinformatics*, vol. 12, no. 1, p. 356, 2011.

[6] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. E. Nelson, "Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC bioinformatics*, vol. 13, no. 1, p. 42, 2012.

[7] Y. Yang, J. Y. Choi, K. Choi, M. Pierce, D. Gannon, and S. Kim, "Biovlab-microarray: Microarray data analysis in virtual environment," in *eScience, 2008. eScience'08. IEEE Fourth International Conference on.* IEEE, 2008, pp. 159–165.

[8] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, "Galaxy cloudman: delivering cloud compute clusters," *BMC bioinformatics*, vol. 11, no. Suppl 12, p. S4, 2010.

[9] L. Jourdren, M. Bernard, M.-A. Dillies, and S. Le Crom, "Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses," *Bioinformatics*, vol. 28, no. 11, pp. 1542–1543, 2012.

[10] [Online]. Available: http://www.genomespace.org

[11] J. T. Dudley, Y. Pouliot, R. Chen, A. A. Morgan, and A. J. Butte, "Translational bioinformatics in the cloud: an affordable alternative," *Genome medicine*, vol. 2, no. 8, p. 51, 2010.

[12] W. Dennis, K. Parul, F. Vincent, P. Rimma, P. Prasad, and T. Peter, "Cloud computing for comparative genomics," *BMC Bioinformatics*, vol. 11, no. 259, pp. 1–12, 2010.

[13] P. Kudtarkar, T. F. DeLuca, V. A. Fusaro, P. J. Tonellato, and D. P. Wall, "Cost-effective cloud computing: a case study using the comparative genomics tool, roundup," *Evolutionary bioinformatics online*, vol. 6, p. 197, 2010.

[14] J. Wilkening, A. Wilke, N. Desai, and F. Meyer, "Using clouds for metagenomics: a case study," in *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on.* IEEE, 2009, pp. 1–6.

[15] T. Gunarathne, T.-L. Wu, J. Y. Choi, S.-H. Bae, and J. Qiu, "Cloud computing paradigms for pleasingly parallel biomedical applications," *Concurrency and Computation: Practice and Experience*, vol. 23, no. 17, pp. 2338–2354, 2011.

[16] R. C. Taylor, "An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics," *BMC bioinformatics*, vol. 11, no. Suppl 12, p. S1, 2010.

[17] A. Matsunaga, M. Tsugawa, and J. Fortes, "Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications," in *eScience, 2008. eScience'08. IEEE Fourth International Conference on.* IEEE, 2008, pp. 222–229.

Consensus in complex networks with state-dependent interactions

Miroslav Mirchev, Aleksandra Bogojeska, Igor Mishkovski and Ljupco Kocarev

Facuty of Computer Sciences and Engineering,
Ss. Cyril and Methodius University in Skopje,
Skopje, R. Macedonia
E-mail:{miroslav.mirchev, aleksandra.bogojeska, igor.mishkovski, ljupco.kocarev}@finki.ukim.mk

***Abstract*—Real networks are characterized by links having properties depending on the environment, nodes' position and state, etc. We study what are the effect of having state-dependent links having a form motivated by the well known Hebb's rule that mimics the plasticity of synapses in the context of the network consensus problem. In addition to our previous research we examine the topology effects on the dynamical behavior by using random and scale-free networks. We also examine other types of dynamical links with similar form.**

I. Introduction

In real networks the interactions, typically referred as links, of the various entities, often called nodes, form complex structures whose properties often can not be immediately seen by simple visualizations. The complexity in these networks can come from the intrinsic properties of the nodes and the way their internal state evolves in times, or from the patterns in which the nodes are interconnected. Both types of complexity have been studied for long time now, and the improvements of experimental equipment, increased computing power and available real data have significantly boosted the research of complex networks resulting in the development of many new models, methods and tools. Two popular review articles of the studies of the topologies of complex networks are [1] and [2], while relatively recently a book aiming to be an introduction to the science of networks has been published by Mark Newman [3]. A well known book by Strogatz is [4] focusing on systems with nonlinear dynamics and chaos with plenty of application examples in various fields like biology, chemistry, physics and engineering.

In the studies of complex networks it is often assumed that links are uniform, however, in real networks typically each link has some unique properties which can be represented by non-uniformly weighted links. We consider dynamical processes in networks of nodes connected with links dependent on the nodes' state, because there are various forms of state-dependent networks in biological, physical, technological and social systems. Hence, state-dependent networks have been studied in many different fields [5]–[9]. As an example, the properties of the wireless links in wireless networks are dependent on the nodes' distance in addition to the environmental conditions [5]. A small example sensor network is shown in Fig. 1 where the octagons are represent simple sensors that transmit the information they gather to the more sophisticated nodes that reroute data and communicate at larger distances, which are represented by squares. The network can also be flat with all nodes (being of the same type) exchanging information with all their neighbors. Another example are neural networks where synaptic efficacy changes according to previous activities of neural cells.

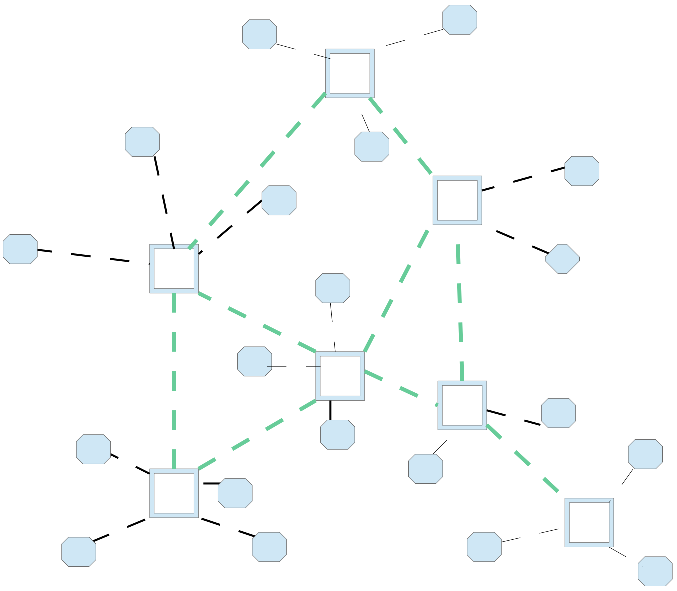


Fig. 1: A small example sensor network where the octagons represent simple sensors and the squares more sophisticated sensor nodes that can communicate at larger distances and process information.

Inspired by the adaptiveness of synapses in neural networks we examine what are the effects of using several types of dynamical weighted links motivated by the Hebb's rule on the process of reaching consensus in a network. The consensus problem is a process of collaboration and exchange of information among networked nodes to eventually agree on a certain value for a given quantity of interest [10], [11], ex. in sensor networks are the measurements and the average measured value. The Hebb's rule is a simplified demonstration of the change of synaptic efficacy in cells depending on their previous stimuluses [12], and one typical description of this rule is to assign weights $w_{ij}(x_i, x_j) = x_i x_j$ to all pairs of connected cells $(i,\ j)$, with x_i being the corresponding stimulus.

In a previous paper [13], we studied consensus in Erdős-Rényi (ER) random networks with links motivated by an *inverted-positive* Hebb's rule. Here we additionally consider another similar form called *positive* Hebb's rule and furthermore, we examine what are the effects of the topology on the dynamics by using the Barabási-Albert (BA) model to generate scale-free networks.

The rest of the paper is structured in the following way. In Section II we define the problem of consensus in networks with two types of state-dependent links. Section III is devoted to the description of the underlying network topologies considered in our analyses. In Section IV we present some numerical results, while with Section V we conclude the paper.

II. Consensus in state-dependent networks

We consider a network of N nodes with information states denoted with a vector $\mathbf{x} = [x_1 \ldots x_N]$. The problem of consensus in state-dependent complex networks can be formulated as in [13],

$$\dot{x}_i = \sum_{j=1}^{n} w_{ij}(x_i, x_j)(x_j - x_i),\ i = 1 \ldots N, \tag{1}$$

where $w_{ij}(x_i, x_j)$ is a state-dependent link between each pair of nodes (i, j).

The network's topology can be represented with a binary adjacency matrix $\mathbf{A} = [a_{ij}]_{N\times N}$, where $a_{ij} = 1$ if there is a link between the nodes pair (i, j) and $a_{ij} = 0$ indicates that the nodes are not connected. The characteristics of the state-dependent links can be then represented with a function $g(\mathbf{x}_i, \mathbf{x}_j)$, $g : R^D \times R^D \to R$ and the instantaneous topology of the entire network with a weighted adjacency matrix $\mathbf{W}(\mathbf{x}) = [w_{ij}(\mathbf{x}_i, \mathbf{x}_j)]$, $w_{ij}(\mathbf{x}_i, \mathbf{x}_j) = a_{ij} g(\mathbf{x}_i, \mathbf{x}_j)$. For simplicity we assume that the interactions are symmetrical $a_{ij} = a_{ji}$, so the network is undirected. The dynamical links can be represented in the following general form

$$w_{ij}(x_i, x_j) = \begin{cases} g(\mathbf{x}_i, \mathbf{x}_j) & \text{if } a_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

If system (1) is connected and undirected, all the node's states eventually converge to one consensus value $\bar{x} = (1/N)\sum_{i=1}^{N} x(0)$, which is the system's equilibrium point [10], [11]. An interesting property is the time it takes for the nodes to achieve consensus, and one known such quantifier of the convergence rate is the second smallest eigenvalue of the network's Laplacian matrix λ_2. Hence, we consider several particular types of dynamical links that could increase λ_2 as well as the consensus convergence rate.

In our previous paper [13] we examined the problem of consensus in Erdős-Rényi (ER) random networks with links motivated by an *inverted-positive* Hebb's rule

$$g^{IP\ Hebb}(x_i^{(1)}, x_j^{(1)}) = (X_O - x_i^{(1)})(X_O - x_j^{(1)}) \tag{3}$$

where an offset X_O is added, which can guarantee that $g^{Hebb}(x_i^{(1)}, x_j^{(1)}) > 0, \forall i, j$, if $X_O > x_i^{(1)}(t)$, $\forall i, t$ or $X_O < x_i^{(1)}(t)$, $\forall i, t$. As the states are *inverted* (taken with negative signs) and the weight function is *positive* we call it *inverted-positive* Hebb's function $g^{IP\ Hebb}(x_i^{(1)}, x_j^{(1)})$.

Here we additionally consider another similar form called *positive* Hebb's rule

$$g^{P\ Hebb}(x_i^{(1)}, x_j^{(1)}) = (X_O + x_i^{(1)})(X_O + x_j^{(1)}), \tag{4}$$

where $X_O < -x_i^{(1)}(t)$, $\forall i, t$ or $X_O > -x_i^{(1)}(t)$, $\forall i, t$.

III. Network topology

All real networks are characterized by a certain pattern of interconnection among the nodes. One property of the interconnections of real networks is that it is typically random. Another property in some types of scale-free networks is the presence of a few nodes which are highly interconnected while the rest of the nodes have a few links, mostly toward the highly interconnected nodes. For example in sensor networks, depending on the application and the characteristics of the sensors one type of networks have nodes that are scattered randomly, but there are also types of networks where there are a few highly interconnected nodes. Similar types of interconnection patterns can be also found in other kinds of real networks. Therefore, we consider two types of topology, random networks generated using the Erdős-Rényi (ER) model and scale-free networks generated according to the Barabási-Albert (BA) model.

A. Erdős-Rényi model

In the ER model [14] a network of N nodes is represented as a graph $G(N, p)$, with p being the probability that a link exists among each pair of nodes. A network is generated using the ER model as a random realization with given values of N and p. Then, it is checked whether the generated network is connected by checking if the second smallest eigenvalue of the network's Laplacian matrix is larger than zero ($\lambda_2 > 0$), otherwise the procedure is repeated until a connected network is obtained. By a simple reasoning it can be concluded that eventually the network should have approximately $L = N(N-1)p/2$ number of undirected links, with the exact number depending on the stochastic realization.

B. Barabási-Albert

The Barabási-Albert (BA) model [15] was developed to incorporate the growth of networks and the tendency of preferential attachment. Unlike the ER model where the number of nodes is fixed in the BA model the network gradually grows starting from a small seed network. As each new node is added links are formed from that node to the existing network in a way that there is a higher probability for a link to be added toward a node that already has a large number of links, a mechanism known as preferential attachment. If L_N links are added with each new node the number of links in the end network is about $L = NL_N$. To keep the networks generated with the two models comparable we use networks with approximately equal number of links $N(N-1)p/2$=NL_N, therefore, in the ER model we set $p = 0.01$ and in the BA model we add $L_N = (N-1)p/2 \approx 5$ links per each node.

IV. Results

We performed numerical simulations of consensus processes in random networks generated using the Erdős-Rényi (ER) model and scale-free networks generated according to the Barabási-Albert (BA) model. The initial states of the nodes were drawn uniformly random from a certain interval (a,b), $a,b \in R$ and the offset X_O was chosen to be larger than the maximal absolute value of the initial states $X_O > |x_i|, \forall i$, although, to keep the weights positive it is enough that $X_O > x_i(0), \forall i$. To keep the setting simple we used $X_O = 2$ and a distribution of initial states around the zero value $(-2,2)$. Simulations of the consensus problem were run on networks of $N = 1000$ nodes with dynamical links following the *inverted-positive* Hebb's rule $g^{IP\ Hebb}$ given by Eq. (3), the *positive* Hebb's rule $g^{P\ Hebb}$ given by Eq. (4) as well as static networks with unitary weights for comparison.

In Fig. 2 is shown the evolution in time of the states of all nodes in a scale-free network and it can be immediately concluded that the dynamical links increase the consensus convergence rate. It can be also observed that the inverted-positive Hebb's rule brings the node's with negative initial state values faster toward the mean of the initial states Ω, while with the positive Hebb's rule the node's with positive initial state values reach Ω faster. It should be noted that the added offset X_O is a translation of coordinates, thus, using a positive Hebb's rule with $X_O = 2$ and $x_i(0) \in [-2,2]$ is equivalent to having $X_O = 0$ and $x_i(0) \in [0,4]$. Similarly, in this setting the inverted-positive Hebb's rule is equivalent to having $X_O = 0$ and $x_i(0) \in [-4,0]$.

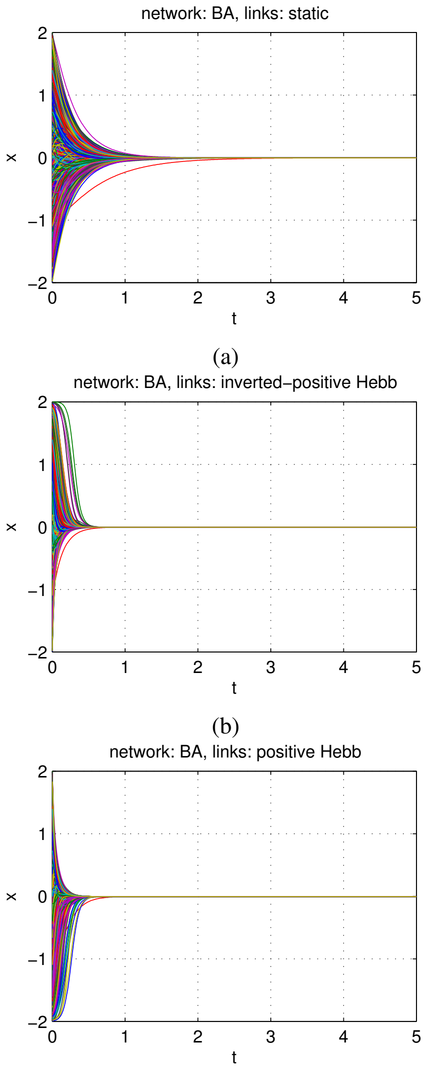
In Fig. 3a we also examine the mean-squared-error (MSE) that indicates the distance of the node's states from the mean of the initial states, which is the expected consensus value. The obtained MSEs confirm that the inverted-positive and the positive Hebb's rules help in achieving network coerence more rapidly. Moreover, it reveals that the convergence is faster in scale-free networks generated with the BA model as it is expected. The convergence rate with the two dynamical link types is similar with very small differences in the initial period.

To further characterize the convergence rate we observe the change in time of the second smallest eigenvalue λ_2 of the network's Laplacian matrix. In case of static networks λ_2 is constant and $\lambda_2 = 0.9$ for the ER model and $\lambda_2 = 1.7$ for the BA model. On the other hand, as shown in Fig. 3b for state-dependent networks with links following the inverted-positive and positive Hebb's rules, λ_2 rises significantly during the initial stage and asymptotically converges to a certain value as consensus is achieved.

V. Conclusion

In this paper we have further studied the effects of the state-dependence of the interconnections on the process of consensus. Moreover, the effects of the topology in this type of state-dependent networks was examined using random and scale-free networks.

It has been shown that the two presented dynamical links provide network convergence to the same average information value, with small differences in the convergence process. As expected the presence of highly interconnected nodes increases the convergence rate in these state-dependent networks, hence, one approach to increase the convergence of consensus in random networks is to introduce a two-level network where the higher level has a scale-free topology.



(a)

(b)

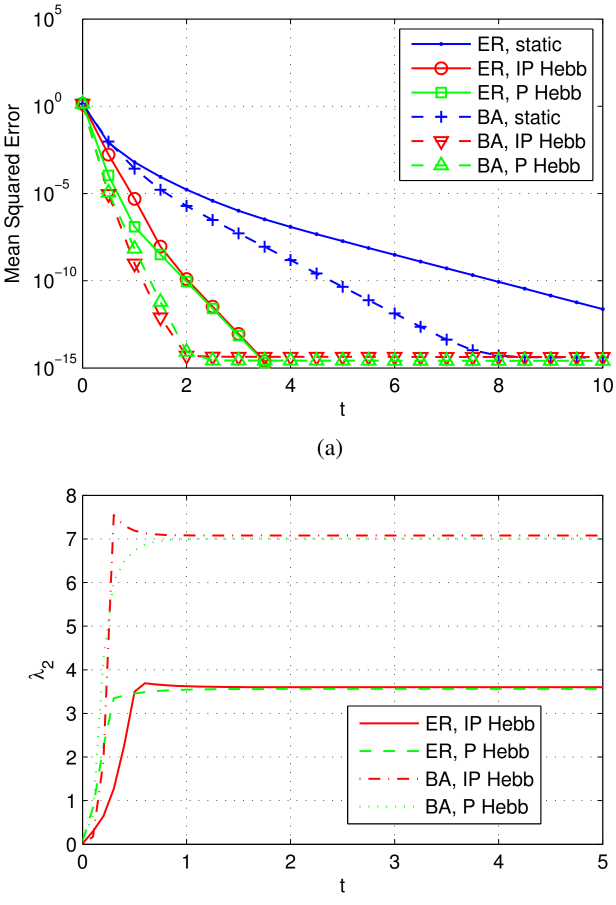
(c)

Fig. 2: A consensus process in networks of $N = 1000$ nodes with initial states in $x_i \in [-2,\ 2]$ for different types of links as indicated in the titles in networks generated using the BA model with $L_N = 5$.

Acknowledgement

The authors thank the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in

Skopje, under the MultiNets ("Modelling and Analysis of Dynamical Processes in Composite and Multiplex Networks") project for financial support.



(a)

(b)

Fig. 3: (a) The Mean squared error and (b) the second smallest eigenvalue λ_2, during a consensus process in networks of $N = 1000$ nodes generated with the ER and BA models, with several types of links and node's initial states $x_i \in [-2, 2]$. The second smallest eigenvalues for static networks are $\lambda_2 = 0.9$ (ER) and $\lambda_2 = 1.7$ (BA).

References

[1] M.E.J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp.167–256, 2003.

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[3] M.E.J. Newman, "Networks: An Introduction", Oxford University Press, Oxford, UK, 2010.

[4] S.H. Strogatz, "Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering," *Addison-Wesley studies in nonlinearity*, Westview Press, 1994.

[5] Y. Kim and M. Mesbahi, "On Maximizing the Second Smallest Eigenvalue of a State-Dependent Graph Laplacian," *IEEE Trans. Autom. Control*, vol. 51, no. 1, pp. 116–120, 2006.

[6] D.D. Siljak, "Dynamic graphs," *Nonlinear Anal.: Hybrid Syst.*, vol. 2, pp. 544–567, 2008.

[7] C. Zhou and J. Kurths, "Dynamical weights and enhanced synchronization in adaptive complex networks," *Phys. Rev. Lett.*, vol. 96, no. 16, p. 164102, 2006.

[8] P. De Lellis, M. di Bernardo, F. Garofalo, and D. Liuzza, "Analysis and stability of consensus in networked control systems," *Applied Mathematics and Computation*, vol. 217, no. 3, pp. 988–1000, 2010.

[9] H. Su, G. Chen, X. Wan,g and Z. Lin, "Adaptive second-order consensus of networked mobile agents with nonlinear dynamics," *Automatica*, vol. 47, no. 2, pp. 368–375, 2011.

[10] R. Olfati-Saber and R.M. Murray, "Consensus protocols for networks of dynamic agents," in *Proc. 2003 Am. Control Conf.*, vol. 2, pp. 951–956, 2003.

[11] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[12] O. Paulsen and T. J. Sejnowski, "Natural patterns of activity and long-term synaptic plasticity," *Current opinion in neurobiology*, vol. 10, no. 2, pp. 172–179, 2000.

[13] A. Bogojeska, M. Mirchev, I. Mishkovski, and L. Kocarev, "Synchronization and consensus in state-dependent networks," *IEEE Trans. Circuits and Systems I*, 2013 (published online).

[14] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[15] Albert-László Barabási and Réka Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

Methods for Designing Test Cases for Web Applications

Bisera Dugalikj
Faculty of Computer Science and Engineering
Skopje, Macedonia
e-mail: bisera_dugalic@yahoo.com

Anastas Mishev
Faculty of Computer Science and Engineering
Skopje, Macedonia
e-mail: anastas.mishev@finki.ukim.mk

***Abstract*— this research is about defining and comparing several approaches regarding the process of designing test cases. We focus on approaches that design test cases that are meant to examine the functionality and how the software behaves when it is used without having an insight into the details of underlying code and internal implementation. This is defined as behavioral or black-box testing and according to its definition the tester is only concerned with what the software is supposed to do, not in the way how the operation is done. The starting point in this kind of testing is for the tester to study the software documentation in detail particularly examining the part that is defined in the requirement technique as basis for defining the test cases. There are many different approaches regarding the requirement techniques depending on their focus like specification-based testing, use case testing, sequence diagram testing, equivalence partitioning, boundary value analysis, state transition tables, fuzz testing, model-based testing, exploratory testing, etc. Since the approach in making the test cases depends on the requirement technique that is selected to be performed in the process, in this research we compare several approaches in order to define their similarities and differences and to discover the cases when we get best results with each different approach. In the end, we define our approach that we intend to use for automatic test case generation.**

Keywords—specification-based testing; use-case testing; sequence diagram testing; equivalence partitioning; model-based testing;

I. Introduction

The importance of software testing as a field of study is constantly increasing as a result to the greater use of software products in ordinary life. The continuous expansion of cloud computing, big data, and social interaction bring new approaches to developing future software products for everyday use. Most recent applications have settings that include huge number of options available for configuration. For example, a recent version of Apache web-server has 172 options that can be configured by users. Most of these configurations are binary but some have three or more settings. It has a total of 1.8×10^{55} different testing states for the system only regarding the settings without considering other system variability that should be tested like user inputs, different combination of operations, etc. Entirely testing a program in some similar cases and same in others that are a lot more simple does not only seem to be time consuming but is also impossible [1]. Even the simplest applications that are created have a lot of options and states that should be tested before the product is released to the customers. Applications like Paint that are plain and used for a long time can also be impossible to test completely if we consider the large number of input values available. Any of the test conditions that are not taken into consideration because the tester would like to save time or the test condition seems to be unnecessary or redundant means that the product will not be completely tested [2].

In this research we focus on several black box testing techniques that we consider to be most effective when it comes to automated test case generation primarily intended for testing enterprise web applications. Black box testing allows us to avoid the concern regarding how the product performs the operations so we can focus on better defining and testing the requirements specification and functional validation of the system. It is concentrated on the end user perspective so the incomplete or not well defined requirements can be noticed and improved at the beginning of the process. With black box testing techniques the tester is concerned with the provided inputs for the application and the expected outputs for stated execution conditions (paths that should be traversed during the execution) and not with what the application does to provide the output or how it happens. Since the tester has no knowledge about the source code and how the program performs the operations, it is less probable to create the tests that match the code's work [3]. When the tester knows the code there is a bigger chance not to include some tests or input and output combination assuming they would pass or because they seem unnecessary regarding how the code works. This approach in testing the applications has several advantages. It offers ways to simplify the testing when it comes to large and complex applications by concentrating on the correct and incorrect input and output combinations obtained. The end user view that the tester has makes it easy to create test cases since there is no need to think about the inner work but only to go through the application like the end user would. Also the process of developing test cases goes a lot faster as a result to following only the user interface paths that take part in the process [4].

II. Designing Good Test Cases

Best way to define test cases is to say that they present a set of inputs, expected outputs and execution paths that are to be followed during the execution of the program intended to test a specific part of the software. A test case for web applications is usually a script containing a list of steps that test the correct behavior and features of the application including an expected final result. The main purpose of test cases is not only to discover defects but also to give information to the tester whether the test passed or failed and the part of the test when it happened. They can help to stop the product from being released prematurely but it can also help finding safe scenarios to use the product so if necessary the product can work in spite of the bugs. This can be very useful when the software has to be released earlier than planned or when one way of doing some task is all the user needs. In some cases it can even be best policy to focus on designing the test cases so that they cover the paths and events that are most likely to be traversed by the user while operating with the web application. Another important thing about test cases is for them to evolve as the program develops. Best way to achieve this is either to change the test cases and make them more complex or to make combinations of two or more test cases designed in the beginning of the process [5].

III. Dominant Black Box Testing Techniques

There are many black box testing techniques that address different aspects concerning the software. Here we present some of them that we intend to use later.

A. *Specification-based testing*

With this technique, test case design is mainly based on the specification, user manual or some notes that define the program's behavior. It allows the test cases to be developed without having seen the product or even without having it developed yet. It can be performed at all levels of testing since it is focused on the requests that are defined in the specification and testing the product against them. This is one of the best ways to test the program regarding how it is expected to work. The main challenge with this method is resolving the right meaning of the specification. In some cases the specification is short and well defined but in other it can be hundreds of pages long or sometimes there can be no documented specification. In some cases large part of the specifications is dedicated to describing some part of the system and it can misdirect the tester to consider that the part is more significant and should be better tested than others. This part of the work is very demanding and in order to make good test cases it is important to make logical links among the requirements, check them for ambiguous meaning, and analyze all the details in order to get a clear picture of how the product is supposed to work.

In order to efficiently automate the test generation in this case the developers need formal specification. Test cases are generated in order to test each of the requirements or test criterion. Test criterion usually addresses some perception of the coverage but it can also be based on a fault domain [6] [7].

B. *Use case and sequence diagram testing*

Use cases are used to describe the functional requirements of the software in detail. They present a sequence of actions the program executes to get a result expected by the user. The Unified Modeling Language (UML) helps to improve the presentation the use cases using diagrams that show the relationships among them and their structure. Use cases are also used in some other methods as a source of information for deriving test requirements that help identify functional test requirements. Use cases define the user's perspective of the program and this testing method is focused on how the user interacts with the software and is best way to test if the product meets the needs of the final user. It tests the structure (relationships) of the use case diagram and it assures the testing of the minimal but essential elements of this diagram. With this approach it is expected to detect errors that are different from the ones identified when use cases are used in functional testing. It is expected to help assess the quality of functional test cases derived from use cases, and test the use case specification itself.

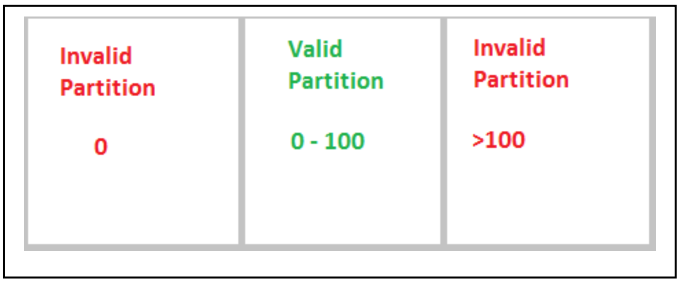
With UML the use cases are graphically represented in the use case diagram but this is not enough to describe the steps of a use case for automation. Testers usually use textual specifications to complement the use case diagram and these specifications use informal textual notation which is adequate to describe the use cases so there is no need for formal specification to define use case descriptions [8] [9].

Sequence diagrams describe the paths the user follows when using the program. It presents the units (methods) that are traversed and the order in which it happens. They are very useful in presenting the dynamic behavior of the system, interfaces with subsystems, as well as some of the flows of the use cases. Same as with use cases, testers in order to avoid formal specification, write informal textual specification that describes the diagram and is adequate for further work especially when it comes to generating test cases with automated test tools [10] [11].

C. *Equivalence partitioning*

This is software testing method that creates test cases based on data partitions that are made by dividing the input and/or output data of a software units. They are designed to cover each partition at least once. The idea of this technique is to reduce the time required for testing the software through reduction of the total number of developed test cases so they can find classes of errors not only individual cases. The partitions are usually defined regarding the specification and the data used with consideration to how it influences the processing of the units. In most cases the equivalence partitioning is practiced to the inputs of the software units but when needed, it can also be done to the outputs. Each partition should have a set of values that are supposed to be equivalent to each other and should be managed by the unit the same manner. Fig. 1 shows a simple case of equivalence partitions. When defining the partitions it is important make correct grouping of the values in order to avoid bigger mistakes later. In the process of selecting values from the partitions it is best to choose more than one value from same partitions especially when it comes to invalid value partitions in order to check not only the system but also the process of partitioning.

The valid input data from the same partition is supposed to go through the same process while the invalid data can go through various processes and should be addressed with caution. With this testing technique, an ideal test case finds a class of errors that usually requires many other (regular) test cases to be executed before the error is found [12] [13].



Invalid Partition	Valid Partition	Invalid Partition
0	0 - 100	>100

Fig. 1. Example: Equivalence partitions with boundary values.

D. Boundary value analysis

This software testing technique is committed to design tests that must include the boundary values of an input domain rather than other acceptable values. The tester defines the boundary values among the set of inputs that are defined to test the software. This is a sort of step two after equivalence partitioning. The inputs that are part of same equivalence class (partition) define the basic set of values for testing. Considering these sets and the inputs they contain, the tester defines the boundary values among them in cases when there is some sort of order among the values. In Fig. 1 there is an example of partitions that present the inputs. With this method the tester includes not only some members to represent each of the partitions but only the boundary values of both valid and invalid partitions.

When it comes to using this method in most cases it presents testing the values on the minimum and maximum edges of the equivalence partitions defined before for the previous method. These values include the input or output ranges of the software component or the internal implementation. They are very often used in designing test cases regarding the fact that the boundaries are natural options for possible errors that result in software faults so it gives the best result when there is some range of values as an input [13] [14].

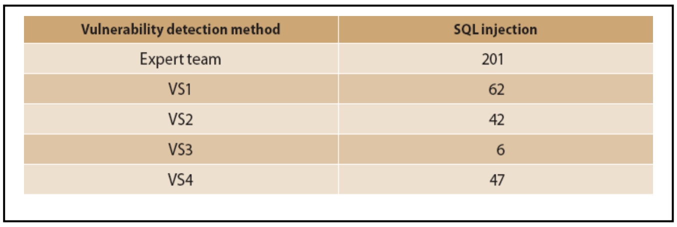
E. Model-based testing

In many cases software testing needs the use of a model to guide the process of test selection and test verification. This approach is similar to specification based testing since the test cases are generated from a model with the difference that the model can represent some aspect of the requirements but it does not need to be the specification. It includes constructing a model of the product and deriving the test cases from the model depending on the notion of coverage. The techniques used are defined by the model they use and how the model is constructed same as its coverage criteria and the analyses used for generating test cases.

The advantages of this technique are the increased productivity with support for visualizing domains and the possibility to start automatically generating test cases at an early stage of the software development, allowing software developer to do coding and testing at the same time [14] [15] [16].

IV. Our Approach in Test Case Design

All currently available automated testing tools implement some of the previously mentioned approaches or similar techniques. According to a recent research made regarding the work of automated testing tools, the results show that their work is not quite satisfactory and could be greatly improved.



Vulnerability detection method	SQL injection
Expert team	201
VS1	62
VS2	42
VS3	6
VS4	47

Fig. 2. Number of errors detected by team of experts and four automated test tools. [17]

In Fig. 2 we can notice that there is a considerable failing in the performance of testing tools compared to the work done by a team of experts. Another worrying fact is that besides the real errors, there is considerable number of false ones. As we can see in Fig. 3, in some cases the number of false errors is higher that the number or real errors discovered by the tools which means that in some cases it will take more time to repair parts of the software that already work instead of focusing on the real failures.

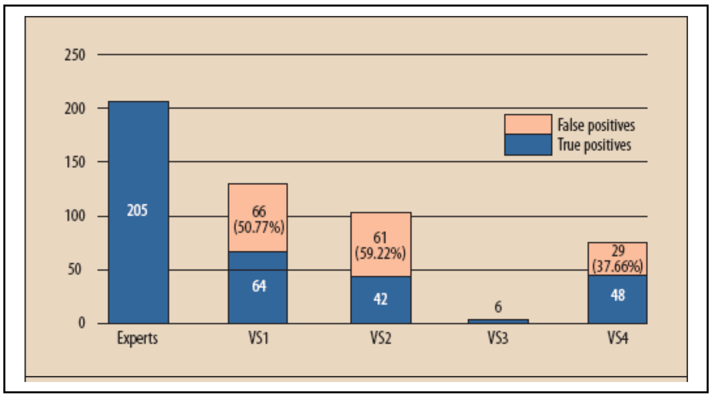


Fig. 3. Number of vulnerabilities detected vs. number of false voulnerabilities. [17]

According to these results the available approaches need to be improved in order to give better results in cases when the testers use automated testing tools.

The low number of real errors detected is not the only concern. As we can see in Fig. 4, it is probable that most testing tools also use similar or maybe even the same approach in generating the test cases since almost all of them find basically the same errors and there is a big number of errors that pass unnoticed. This means that not even using more than one automated testing tool at the same time will bring better results and find more errors compared to what we can get by using only one tool.

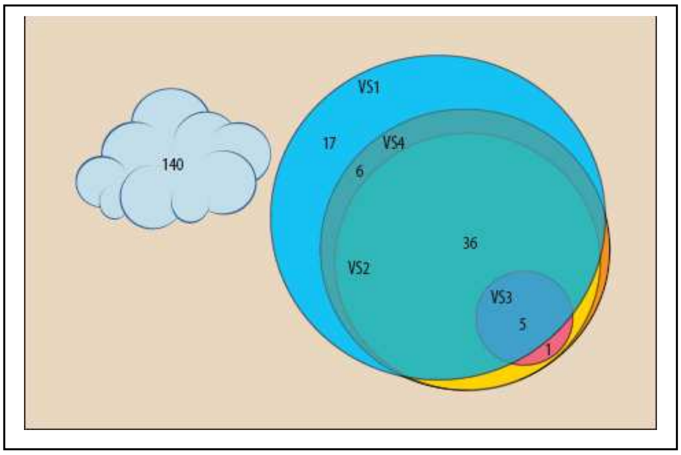


Fig. 4. Overlap among sets of errors detectedby each scanner. (The areas in the circles are proportional to the number of errors detected). [17]

The main problem might be that all available testing techniques were primary intended to be used by expert teams, not to be implemented into testing tools. In our opinion, it is possible to get far better results by combining some of the previously mentioned methods. We present a model that puts together some of the earlier described techniques improved with some ideas that come from testers and are used while performing manual software testing. Since our goal is to test the logic flow of enterprise web applications we intend to base our model on use case testing technique. We intend to cover both options, generating test cases with and without available formal specification. When there is no formal specification of the product (that can often be an issue), we intend to generate test cases according to the available use cases and sequence diagrams. Sequence diagrams are more intended to improve the use cases and to help avoid any ambiguity that might be found in the use cases than to be actually used in the process of test case generation. The number of use cases defines the number of test suites (collections of test cases) and the test cases they contain are developed depending on what is the purpose of each test case. If there is formal specification, the number of test suits is defined by the number of requirements. The test cases in each of the test suits will be developed to test the logic flow but in this part we will also implement some ideas generally from equivalence partitioning, boundary value analysis and fuzz testing. Fuzz testing is intended to check the security problems of the application considering that it provides invalid, unexpected, or random data as an input. We also include some ideas from other black-box techniques that seem to be useful in capturing what we could define as a kind of simulation of human behavior in the process of software testing and some undefined techniques that are done by testers while doing manual testing that could be automated [17].

V. Conclusion

In our opinion, all testing techniques have their specific purpose and in some cases one approach can give much better results that other. Our approach is intended to make a combination of some of these methods so that we can get the benefits of all the approaches we choose to implement. This might not give as good results in some cases as would one technique but in general it is expected to be much easier to use and to cover a bigger part of the testing process.

VI. Future Work

Our future work consists in creating an algorithm for automatic test case generation that will be appropriate for the model we described. It will include parts of some of the earlier mentioned black-box techniques in order to give better results in generating test cases. We also intend to add some trivial steps to the algorithm that might not be included in any of the mentioned methods but any tester that does the testing manually would include them.

Our main goal with this algorithm that is generally intended to be implemented in some tool that automatically generates test cases is to create test cases that cover the logical flow of the application and we intend to avoid the cases of false positive results when it comes to errors.

References

[1] C. Yilmaz, S. Fouche, M. B. Cohen, A. Porter, G. Demiroz, and U. Koc, "Moving forward with combinatoral interaction testing", IEEE Computer society R. Bryce, R. Kuhn, February 2014, pp. 37-45.

[2] R. Patton, "Software testing", November 2000, pp. 63-89.

[3] S. Nidhra and J. Dondeti, "Black Box and White Box Testing Techniques – A Literature Review", International Journal of Embedded Systems and Applications (IJESA) Vol.2, No.2, June 2012.

[4] H. V. Kantamneni, S. R. Pillai, and Y. K. Malaiya, "Structurally Guided Black Box Testing", 2002.

[5] C. Kaner, "What Is a Good Test Case?", STAR East, May 2003.

[6] T. Kahsai, M. Roggenbach, and B.H. Schlingloffy, "Specification-based Testing for Software Product Lines", Humboldt University Berlin Fraunhofer, 2008.

[7] A. D. Brucker, M. P. Krieger, D. Longuet, and B. Wolff, "A Specification-based Test Case Generation Method for UML/OCL", 2011.

[8] S. K. Swain, D. P. Mohapatra, and R. Mall, "Test Case Generation Based on Use case and Sequence Diagram", Int.J. of Software Engineering, IJSE Vol.3 No.2, July 2010.

[9] A. Carniello, M. Jino, M. L. Chaim, "Structural Testing with Use Cases", 2005.

[10] E. G. Cartaxo, F. G. O. Neto, and P. D. L. Machado, "Test Case Generation by means of UML Sequence Diagrams and Labeled Transition Systems", 2007.

[11] A. Rountev, S. Kagan, and J. Sawin, "Coverage Criteria for Testing of Object Interactions in Sequence Diagrams", 2005.

[12] S. C. Reid, "An Empirical Analysis of Equivalence Partitioning, Boundary Value Analysis and Random Testing", 1997.

[13] V. Arnicane, "Complexity of Equivalence Class and Boundary Value Testing Methods", Scientific Papers , University of Latvia, Vol. 751, Computer Science and Information Technologies pp. 80-101, 2009.

[14] A. Pretschner, "Model-Based Testing in Practice", J.S. Fitzgerald, I.J. Hayes, and A. Tarlecki (Eds.): FM 2005, LNCS 3582, pp. 537–541, 2005.

[15] I. K. El-Far, and J. A. Whittaker, "Model-based Software Testing", Encyclopedia on Software Engineering (edited by J.J. Marciniak), Wiley, 2001.

[16] M. Utting, A. Pretschner, and B. Legeard, "A Taxonomy of Model-based Testing", April 2006.

[17] N. Antunes, and M. Vieira, "Penetration Testing for Web Services", IEEE Computer society R. Bryce, R. Kuhn, February 2014, pp. 30-36.

On the maximum utilization of the mother steel plates for rectangular steel plates orders

Lasko Kasapinov
Open Mind Solutions LTD
St. Kliment Ohridski 20A-3
Skopje, Macedonia
lasko.kasapinov@gmail.com,

Mile Jovanov and Dimitar Trajanov
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
{mile.jovanov, dimitar.trajanov}@finki.ukim.mk

Abstract – **The paper introduces a problem that arises from production process in steel industry. This problem involves matching orders against inventory slabs and it is a combinatorial NP-complete problem. Further in this paper, we discuss the different approaches on solving the problem presented and used through the years. We also present our implemented solution for one variation of the problem. At the end, we present the achieved results, the savings made in the last years in the factory, after implementation of the solution.**

Keywords— steel industry, production process, slabs, mother steel plates, orders, algorithms

I. Introduction

When small items are being cut out from large objects, two problems arise. The first one is the assortment problem addressing the issue of choosing proper dimensions for the large objects. The second one is the trim loss problem addressing the issue of how to cut out the small items from the given large objects in such a way that waste material will be minimized (Hinxman, 1979). The combination of the assortment problem and the trim loss problem is known as the cutting stock problem (CSP).

The CSP arises in many industries where large materials must be cut into smaller pieces. This paper is concentrated on production of rectangular steel plates which are cut from large steel plate produced from slabs by rolling. The slabs have cuboids shape with different dimensions and are obtained by casting the molten steel. The steel plate has the same chemical structure like the used slab it is produced from, and the plate dimensions are adequate to the order the steel plate is produced for.

The process of production planning usually begins with the list of orders that need to be met with the appropriate priority, and list of slabs on the other side. The solution to the CSP is a cutting plan which combines the ordered steel pieces for maximum utilization of the material from slabs. Resolving this issue is important because minimal improvements in the manufacturing process can result in huge savings.

The general problems associated with the standard cutting of small items are known to be NP-complete problems. In many cases this kind of problems can be modeled using mathematical programming and solution can be found using approximate methods or some heuristics. The goal is designing a plan for one-dimensional cutting of material in many pieces with minimal overall loss while various requirements that may occur in practice will be satisfied.

This paper continues with a literature study covering few different formulations and solution methods of the CSP. Furthermore, we present a formulation for the specific CSP in the steel industries and an approach and algorithm for solving it. Finally, we present the achieved results.

II. CSP's concepts

There are several known attempts to resolve the problem of cutting the materials into products in theory. The beginnings and the first definitions are introduced by Kantorovich since 1960. Scientific research has led to several formulations and solution. The most famous formulation was composed by Dyckhoff [8] in 1990. In his work he has defined a schema for classification of the problems based on four characteristics: size, type of task, the semi-product assortment (raw material) and product range.

The CSP is a kind of matching problem on a sparse graph, allocating set of materials to a given set of orders. A bipartite graph is constructed with two sets of nodes, one for orders and the other for the inventory slabs. The edges in this graph represent the potential matches of orders against slabs according to the production constraints (Fig. 1). The assigning restrictions convert this graph to a bipartite sparse graph.

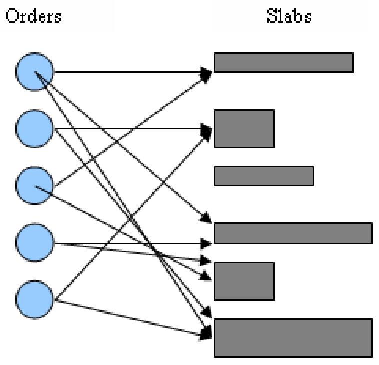


Figure 1. Potential matches between orders and slabs

Each order contains information on the required quality and dimensions of the plates. Each order limits its feasible slabs by means of quality and a slab can be used for production of several orders. But, there are many eligible slabs with different geometries for each order. So, for the given weight of mother-plate, the length of slab is strongly dependent on the width of the slab as it is shown in Fig.2. Also, there are many production constraints that make the problem hard to solve, and the solution to such a problem depends on the size and assortment of the products, and the technology applied in the production process.

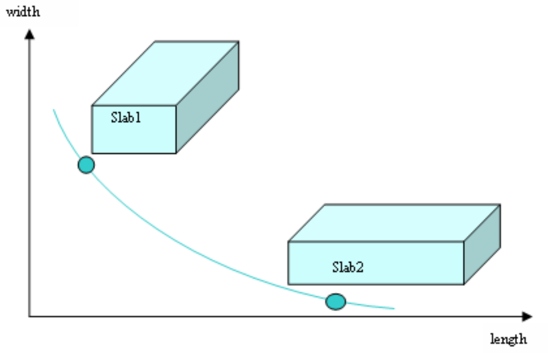


Figure 2. Eligible slabs with different geometries for each proposal

The optimal cutting plan design (examples given in Fig. 3) deals with the following issues while all the production constraints are satisfied:

1. minimal waste material;
2. reduced number of created slabs without orders;
3. improved capacity utilization (slabs strive to be as much as possible within the limits).

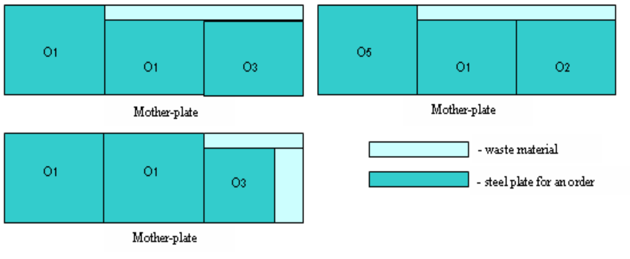


Figure 3. Cutting plan examples

III. Review of the Related Work

There are few different approaches and solution methods for solving CSP in the literature. Generally, the algorithmic methods guarantee the optimal solution for the problem to be found. Gilmore and Gomory [2] have obtained a solution by Linear Programming (LP). They have suggested "column generation" method which has been showed excellent results over a small number of subjects. The drawback of their approach is high running times on larger instances because of computational complexity of the solution. Furthermore, the process of production steel plates can obtain wide range of products and many limiting factors which additionally expand the solution space and possible combinations as well as execution time i.e. execution time is significantly longer. Heinz et al. in 2011 [15] have introduced an integer program solver to solve the problem by removing the naturally arising symmetries. Further, Gargani and Refalo [3] have stated that for constraint programming approach is not ideal to use binary variables in the basic linear models because of their restrictions on values. Integer programming solvers can tighten the formulation by adding cutting planes and they use the relaxed optimal solution to guide the search. Therefore the authors, based on the structure of the problem have introduced logical global constraints. Additionally, with their particular strategy for variables and their values, they have defined a particular heuristics for searching the results around specific combination i.e. Large Neighborhood Search (LNS). Moreover, by using certain static constraints in their model, they are trying to avoid creating symmetrical solutions. Finally, in 2007 they have got pretty fast solution. But, this solution is still suitable for smaller instances only, while looking for results in case of larger instances this solution leads to poor performance. For this reason, algorithmic approaches for solving the CSP have been supplemented with heuristics. Heuristic methods generate faster an acceptable solution which is close enough to the known or believed optimal solution. Heuristic methods are also highly domain-dependent, in sense that they use information about the particular problems for which they are developed. Thus they may appear to almost useless on apparently similar problems (Hinxman, 1979). The basics of heuristics' approach are introduced by Haessler [6]. In his papers he has proposed that sequences of setting the product should be produced and should be generated using heuristics, and this would give fast but not exactly the most optimal solution. So, speed is achieved but the proposed solution is not as effective.

There has been also an idea for generation a hybrid solution which combines algorithmic methods with heuristics. In that manner, the researchers have introduced an approach which uses the Simplex method and the process of determining the next cutting pattern is based on a pattern obtained by LP solution complemented with global and logical constraints. The delayed column generation (DCG) technique is introduced with this approach and it provides greater and more efficient propagation in the search of solutions.

There have been meta-heuristic methods also which have an ability of not being trapped into local optima as might happen with traditional heuristics. The solution process is often guided by some lower level heuristic, when using meta-heuristic methods, and finally results in much desired answers with little solving time. Eshghi [25] (2005) has designed an Ant Colony Optimization (ACO) algorithm for solving one-dimensional cutting problem. In this algorithm based on designed probabilistic laws and improvements, the way of cutting large objects to satisfy demand by artificial ants is provided. These Sequential Heuristic Procedures outperform the possibilities of linear programming and can be considered for solving of more complex problems. Using these procedures the decision is built sequentially (one by one) by defining templates until all the necessary requirements are met.

Although meta-heuristic methods give good enough solutions, the solutions obtained by algorithmic approach have higher quality. If we can reduce the execution time of the algorithmic procedures then we will have better solution.

IV. THE IMPLEMENTED SOLUTION

The solution to such a problem depends on the size and diversity of the products and the technology applied in the production process, and therefore we have to develop a special algorithm for solving any specific problem. The algorithm that we have implemented in a local company is developed in Oracle PLSQL programming language and it is integrated in an ERP environment.

The main reason why we decided to use the linear programming approach is the production technology used in the local company. There are several factors that affect the production process in terms of achieving the required mechanical and chemical properties and their appropriate control, proper transformation from raw material to final product and producing the required dimensions of final products. It is not easy to implement all these rules into some of the existing platforms designed for solving this type of problem.

Our approach uses LP procedures complemented with branch and bound technique. So, to reduce the execution time necessary to propagate to solution we used the best practices of the constraint programming community. The algorithm starts with defining the relation between slabs and orders in such manner that provides information about which slab can be used for production of each order according to the rules. This selection prevents the algorithm from irregular combinations because in the next step the algorithm makes all possible combination of previously selected orders for each slab. The process of combining the orders uses the weight of orders to fill the slab until the full capacity of slab is reached or exceeded. The implemented dynamic symmetry-breaking scheme leads to unique and asymmetrical combinations. Additionally, orders are combined starting with the biggest order – a heuristic that provides us with better propagation in searching.

The algorithm uses greedy approach and tries to fill the slab with one order before it starts to combine the next order. With the aim to decrease execution time an implemented heuristics stops the generation of solutions if several best solutions are generated. In the last step of the algorithm, the results are sorted and a combination is selected for any slab ensuring the best overall quality. This approach finds the patterns and possible solutions in a sequential and iterative way, and the execution time is decreased by used heuristics and implemented constraints. Finally, the algorithm mentioned above obtains high quality results within the defined time limit, and provides large savings in the company, which we analyze in the following section.

V. EXPERIMENTS AND RESULTS

In this section we will present the real-life results of the implemented solution. The test data sets were obtained from existing ERP system and they are real-life data. To monitor the performance of the proposed algorithm under various conditions, the test instances are grouped into three different data sets: small data set composed of instances with around 1.500t slabs and almost the same amount of orders, medium data set in which an instance contains slabs with total weight of around 5.000t, and large data set where the instances contain total weight of 10.000t slabs. There were 20 different instances in the small data set, and the medium and large test data sets involved 5 different instances.

The results obtained from computational tests using instances with different size are summarized in Table 1 and graphically presented in Figure 4 to Figure 6.

TABLE I. PERFORMANCE OF ALGORTIHM FOR CSP

	Slabs		Orders		
Data set	**Pcs**	**Weight (t)**	**Pcs**	**Weight (t)**	**Execution Time (s)**
Small data set	300	1490	500	1600	25
Inter-mediate	1000	4890	1500	4700	130
Large data set	2000	9650	3000	9500	360

The data sets specification (number and weight of slabs and orders) are rounded. It is noticeable that our algorithm creates solutions very quickly for smaller instances, and the performance are declining with increasing instances as it is expected. The results show the performance of used heuristics and solutions provided for large test data set are found within a time limit of 10 minutes.

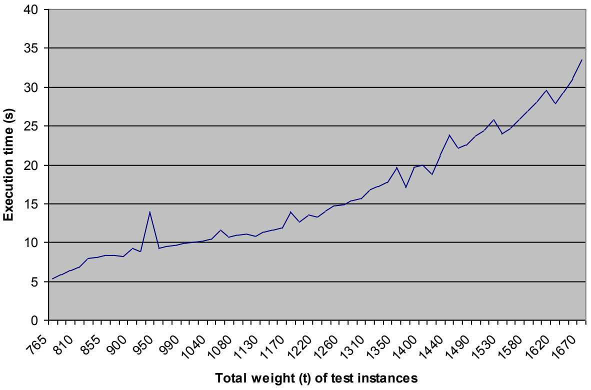


Figure 4. Performance of algorithm with small data set

The test data are intentionally created from real-life data with presented dimension. The instances from small data set represent more then an average daily production, while the instances from large data set represent volume of a two weeks production. And, the results show that managers can make a production plan for the next two weeks within 6 minutes. On the other side, we can notice that the execution time increases as the instances become larger. If the production level is higher than the algorithm will run out the time limit.

TABLE II. THE QUALITY OF SOLUTIONS

	Slabs		Orders		Used Slabs		Satisfied orders		
Data set	Pcs	Weight (t)	Pcs	Weight (t)	Pcs	Weight (t)	Pcs	Weight (t)	Trim loss (%)
Small data set	300	1490	500	1600	220	1150	360	1130	1,74
Inter-mediate	1000	4890	1500	4700	710	3700	1200	3620	2,16
Large data set	2000	9650	3000	9500	1300	7900	2460	7610	3,67

In terms of quality of generated solutions presented in Table 2, there is enormous progress compared to the previous performance of the company although our algorithm does not provide the best solutions due to the implemented heuristics. Thereby improving the quality of solutions, the algorithm provides better utilization of available orders (weight of satisfied orders over weight of orders) and ensures better usage of slabs i.e. less trim loss and large financial savings. There are implicit financial savings also: reduced stock material, faulty-less products and improved efficiencies in production.

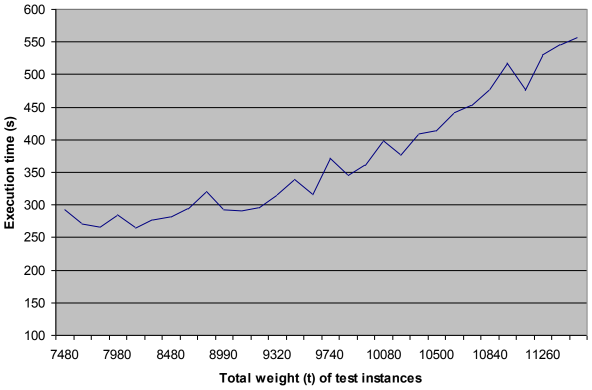


Figure 5. Performance of algorithm with large data set

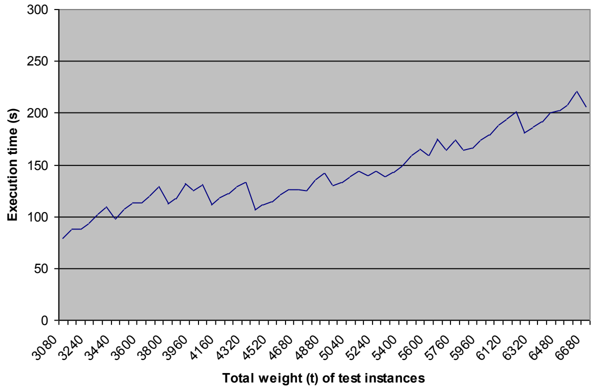


Figure 6. Performance of algorithm with medium-size data set

VI. CONCLUSION

In this paper we presented our work on implementation of an algorithmic solution for the process of production of mother-plates from slabs. The results show the fact that a portion of construction waste can be reduced by a better planning which is done by the program instead of a worker. Using optimization techniques not only reduces waste materials but also improve financial performance of the company by improving the investment level of the stock material. Furthermore, an optimized scheduling of cutting stocks underlines management ability of a company.

ACKNOWLEDGMENT

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] Goutham Dutta, (2001), “A Survey of Mathematical Programming Applications in Integrated Steel Plants”, *Informs Published Online: October 1, 2001*

[2] P.C. Gilmore and R.E. Gomory, (1963), “A linear programming approach to the cutting-stock problem” – Part II, *Operations Research*

[3] Antoine Gargani and Philippe Refalo (2007), “An Efficient Model and Strategy for the Steel Mill Slab Design Problem”, *Principles and Practice of Constraint Programming – CP 2007*

[4] H. Yanagisawa (2007), “The material allocation problem in the steel industry”, *IBM Journal of Research and Development, Volume 51, Number 3/4*

[5] Jayant Kalagnanam (2000), The Surplus Inventory Matching Problem in the Process Industry, *Operations Research, Vol. 48, No. 4*

[6] Robert W. Haessler and Paul E. Sweeney, “Cutting stock problems and solution procedures,” European Journal of Operational Research, Volume 54, pp. 141-150, 1991

[7] Dyckhoff, H., Finke, U., Kruse, H.-J. (1988), "Standard software for cutting stock management", *Essays on Production Theory and Planning.*

[8] Dyckhoff H., “A typology of cutting and packing problems”, *European Journal of Operational Research 44 (1990)*

[9] Eugene J. Zak, “Row and column generation technique for a multistage cutting stock problem”, *Computers & Operations Research 29 (2002)*

[10] Stefan Heinz, Thomas Schlechte, Rudiger Stephan (2010), “Solving Steel Mill Slab Problems with Branch and Price”, *ZIB-Report 09-14 (May 2010)*

[11] L. Fernandez, C. Pola (2008), "Integer Solutions to Cutting Stock Problems", *Dpto. Matematicas, Estadstica y Computacion, Universidad de Cantabria, Santander, Spain*

[12] Lijun Wei, Defu Zhang (2008), "A least wasted first heuristic algorithm for the rectangular packing problem", *Computers & Operations Research*

[13] Ernesto G. Birgin, Rafael D. Lobato, Reinaldo Morabito (2010), "Generating unconstrained two-dimensional non-guillotine cutting patterns by a recursive partitioning algorithm"

[14] Hans Kellerer, Ulrich Pferschy, David Pisinger, (2004) "Knapsack Problem"

[15] Stefan Heinz, Thomas Schlechte, Rudiger Stephan, Michael Winkler, (2011) „Solving steel mill slab design problems", *ZIB-Report 11-38 (September 2011)*

[16] Jayant Kalagnanam, IBM Research Division (1998) "Inventory Matching Problems in Steel Industry"

[17] A.Schrijver, (1986), "Theory of Linear and Integer Programming"

[18] Sang Hwa Song, Sang Min Park, Ho Ki Nam (2006) *"*A Set-Partitioning Apprach to the Inventory Allocation Problem in the Steel Making Industry*", Proceedings of the 7th Asia Pacific Industrial Engineering and Management Systems Conference*

[19] Iiro Harjunkoski, Ignacio E. Grossmann (2001), "A decomposition approach for the scheduling of a steel plant production"

[20] Iiro Harjunkoski, Ignacio E. Grossmann (2001), "A decomposition approach for the scheduling of a steel plant production"

[21] Andrea Bettinelli (2006), "A branch-and-price algorithm for the two-dimensional level strip packing problem", *Universitµa degli Studi di Milano*

[22] Andreas Bortfeldt, Tobias Winter (2009), "A genetic algorithm for the two-dimensional knapsack problem with rectangular pieces*"*

[23] Jakob Puchinger, Günther R. Raid (2005), "A hybrid approach for optimization of one dimensional cutting"

[24] G.Belov, G.Scheithawer (2002), "A cutting plane algorithm for the one-dimensional cutting stock problem with multiple stock lengths",*European Journal of Operational Research 141, 274-294*

[25] Yaodong Cui, Yuli Yang (2006), "A recursive branch-and-bound algorithm for the rectangular guillotine strip packing problem", *European Journal of Operational Research*

[26] K. Eshghi (2005), *"An ACO algorithm for one-dimensional cutting stock problem", Journal of Industrial Engineering International September 2005, Vol.1, No.1*

[27] Harald Dyckhoff, Ute Finke (1992), "Cutting and packing in production and distribution: a typology and bibliography"

[28] Cynthia Barnhart, Ellis L. Johnson, George L. Nemhauser, "Branch and Price: Column Generation for Solving Huge Integer Programs", *Informs Volume 46 Issue 3, May-June 1998, pp. 316-329*

Optical character recognition applied on receipts printed in Macedonian language

Martin Gjoreski, Gorjan Zajkovski, Aleksandar Bogatinov,
Gjorgji Madjarov, Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Skopje, Macedonia

Hristijan Gjoreski
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia

***Abstract*— The paper presents an approach to Optical Character Recognition (OCR) applied on receipts printed in Macedonian language. The OCR engine recognizes the characters of the receipt and extracts some useful information, such as: the name of the market, the names of the products purchased, the prices of the products, the total amount of money spent, and also the date and the time of the purchase. We used the publicly available OCR framework Tesseract, which was trained on pictures of receipts printed in Macedonian language. The results showed that it can recognize the characters with 93% accuracy. Additionally, we used another approach that uses the original Tesseract to extract the features out of the picture and the final classification was performed with k-nearest neighbor's classifier using dynamic time warping as a distance metrics. Even though the accuracy achieved with the modified approach was for 6 percentage points lower than the original approach, it is a proof of concept and we plan to further research it in future publications. The additional analysis of the results showed that the accuracy is higher for the words which are prescribed for each receipt, such as the date and the time of the purchase and the total amount of money spent.**

Keywords—OCR; Receipt digitalization; Tesseract; DTW;

I. Introduction and Related work

Optical Character Recognition (OCR) is conversion of photographed or scanned images, which contain printed or typewritten text, into machine readable characters (text). The basic idea origins since 1929 when the first OCR patent is obtained by Tausheck [1]. It is based on template matching by using optics and mechanics. After the first commercial computer (UNIVAC I) is installed (1951), the era of converting images of text into computer readable text has started. In 1956 the first approach to convert images of text into computer readable text was presented [2]. At that time hardware and software are strong limitations, so the OCR approaches are based on template matching and simple algebraic operations. Since then a lot research has been done on OCR and with the advancement of the technology more complex OCR approaches are developed. Today OCR is done in much more intelligent way, but it also requires more computational power, which can be a problem for smartphone implementations.

OCR can be used in common industries and applications including date tracking on pharmaceutical or food packaging, sorting mail at post offices and other document handling applications, reading serial numbers in automotive or electronics applications, passport processing, secure document processing (checks, financial documents, bills), postal tracking, publishing, consumer goods packaging (batch codes, lot codes, expiration dates), and clinical applications. Also OCR readers and software can be used, as well as smart cameras and vision systems which have additional capabilities like barcode reading and product inspection.

In recent years, numerous OCR-based smartphone applications were also introduced. A successful example application is the Google's Goggles application [3], which has more than 10 million downloads. Beside the OCR functionality it has several others such as: image search, text translation, bar code scanner. Their OCR engine can analyze text in several languages, not including the Macedonian language. Additionally, the implementation of the OCR engine is not on the smartphone itself, but on a server and therefore it requires internet connection in order to perform an OCR action. Recently Google has allowed public and freely available API for their OCR engine [4], which resulted in numerous smartphone OCR-based applications. However they can only be used with internet connection and furthermore, the API does not provide support for the Macedonian language. Finally, there are some examples of OCR-based applications that claim to support the Macedonian language, e.g., Translang [5]. However, none of them supports an OCR for Cyrillic script, which is the official script of the Macedonian language.

In this paper we present an application of OCR on receipts printed in Macedonian language. The next section presents the methodology used for the process of OCR. Then, in the Experimental Results section, the recognition accuracy is presented. Finally, the conclusion and a brief discussion about the approach and the results are given.

II. Methodology

Figure 1 shows the whole process of the OCR. First, the user takes a photo of a receipt that he/she received from a market. Then, the OCR engine recognizes the characters printed on the receipt and therefore extracts some useful information out of the receipt, e.g., the name of the market, the names of the products purchased by the user, the prices of the products, the total amount of money spent, and also the date and the time of the purchase. For the process of the OCR, the open source OCR engine called Tesseract [6] is used.

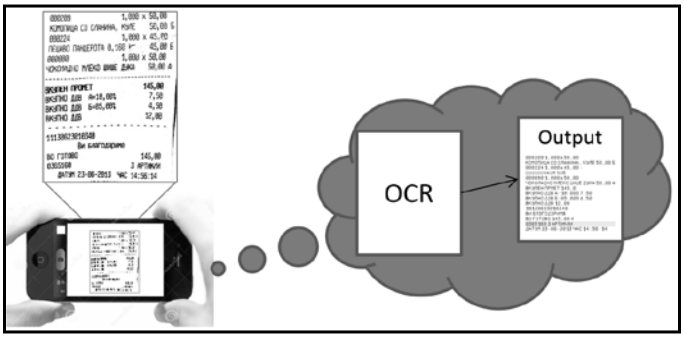


Figure 1. The OCR process applied on a receipt printed in Macedonian language (Cyrillic script).

A. Tesseract

Creating an OCR engine is a challenging research task and requires great knowledge in image processing, feature extraction and machine learning. However, there are several open source projects that provide OCR framework and are widely used in the creation of OCR-related applications. In order not to reinvent the wheel and also to save time for development, in this study we decided to use an OCR framework which is freely available. After studying several frameworks, we decided to use the Tesseract. Tesseract is OCR engine that is developed by HP between 1984 and 1994 to run in a desktop scanner, but it is never used in an HP product [7]. Since then it has a lot of improvements. In 2005 it becomes open source and is managed by Google since then. The last stable version (V3.02) is released in 2012 and V3.03 is expected to be released in 2014. Tesseract is written in C and C++ but it also has Android and iOS wrappers which make it useful for smartphone application.

1) Tesseract Architecture

The first approach that is tested in the process of character recognition is the original Tesseract engine. Tesseract has traditional step-by-step pipeline architecture (shown in Figure 2). First image preprocessing is done with adaptive thresholding where a binary image is produced. Then connected component analysis is done to provide character outlines. Next techniques for character chopping and character association are used to organize the outlines into words. In the end two-pass word recognition is done by using methods of clustering and classification. For the final decision about the recognized word, Tesseract consults with both language dictionary and user defined dictionary. The word with smallest distance is provided as an output. This is just brief overview of the Tesseract architecture, more details can be found in the authors' literature [7].

2) Training Tesseract

For the training phase, Tesseract needs a photograph (tiff or pdf file) of a text written in the same language as the one that it is trying to recognize. For each character from the learning text Tesseract extracts 4 different feature vectors. Then it uses clustering technique to construct a model for each character and those models are later used in the classification phase for decision of which character should be recognized.

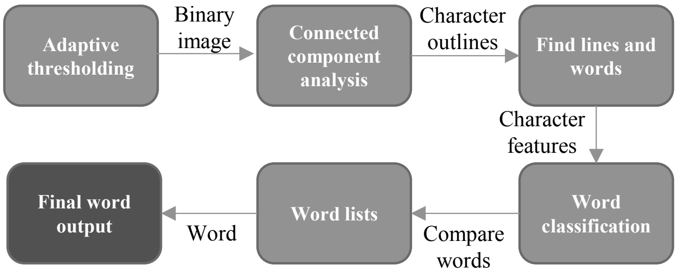


Figure 2. Tesseract OCR engine architecture.

For preparing the training text, several different approaches were tested regarding the font of the training text, the size of the characters in the training text and the content of the training text. Tests were done for each of the three problems. In the first approach the training text was written with a font that was made of quality photographs of single characters. In the second approach the training text was written with a font that is similar to the font of the receipts. For the size of the characters in the training text tests were done with different font sizes starting from 16px to 48px. Regarding the content of the training text two different approaches were tested. With the first approach for each character that the model is trying to recognize there are 10 to 25 different instances with respect to the frequency of the character in the Macedonian language. For example the count of the vowels was 20-25 and the count of the special characters or very infrequent characters such as Н or Z was 10-15. In the second approach the training text was consisted of 1300-1500 random sampled words from different receipts.

The tests showed that the engine is most accurate if the size of the letters in the training text is similar to the text on the photographed receipts. In this case the size that is used is 40px. Also it was concluded that better results can be achieved if the training text is consisted of random sampled words from different receipts the second approach. After all the testing done on Tesseract, the training text that was used for further analysis consisted of 1300-1500 random sampled words from different receipts, it was written with a font similar to the font of the receipts and the size of the characters was 40px.

B. Tesseract-DTW

For the process of character recognition we also tried another approach that uses Dynamic Time Warping (DTW) [8] and K-Nearest Neighbors (KNN) classifier [9]. This approach, Tesseract-DTW (shown in Figure 3), uses the original Tesseract only for feature extraction; the final classification is performed by the KNN classifier using the DTW as a distance metrics. The DTW metric was chosen because the size of each feature vector extracted by the Tesseract varies, and is not the same for each character. Please note that applying a standard classifier such as decision tree, SVM, etc., was not an option because of the varying size of the feature vectors.

1) DTW

DTW also known as dynamic programming matching is a well-known technique to find an optimal alignment between two given sequences [8]. It finds an optimal match between two sequences of feature vectors by allowing stretching and compression of sections of the sequences. DTW first has been used by Sakoe and Chiba [10] to compare different speech patterns in automatic speech recognition. In fields such as data mining and information retrieval, DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data. Also it successfully has been used both for online [11] and offline signature verification [12].

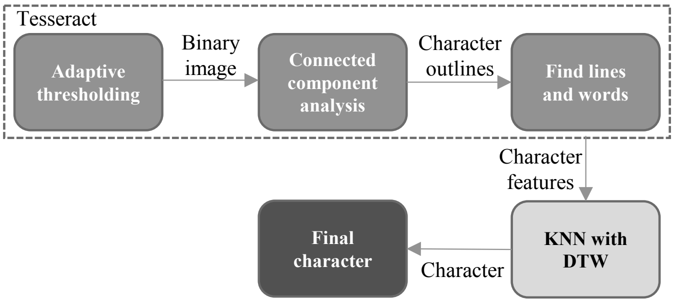


Figure 3. Tesseract-DTW architecture.

2) DTW distance

To calculate the distance between two vectors *X1 = (x11, x12, ..., x1i),* and *X2 = (x21, x22, ..., x2j)*, DTW needs a local cost measure, sometimes also referred to as local distance measure. In this study an Euclidean distance is used as cost measure, see equation (1). By evaluating the local cost measure for each pair of elements of the sequences *X1* and *X2*, cost matrix *M* is calculated, see equation (2). The goal is to find an alignment between *X1* and *X2* having minimal overall cost. For calculating the minimal overall cost three conditions must be satisfied: boundary condition, monotonicity condition and step size condition. The minimal overall cost is the output of the DTW algorithm, shown in equation (4).

$$Cost\ (x1i, x2j) = Euclid\ (x1i, x2j) \qquad (1)$$

$$M[i][j] = Cost\ (x1i, x2j) \qquad (2)$$

$$DTWdist\ (X1, X2) = M_{[1][1]} + S_{min} + M_{[i][j]} \qquad (3)$$

Where, $S_{\min} = \sum (\min (M_{[k+1][t]}, M_{[k][t+1]}, M_{[k+1][t+1]}))$, $k \epsilon \{1, 2, ..., i\text{-}2\}$ and $t \epsilon \{1, 2, ..., j\text{-}2\}$.

3) Evaluating Tesseract-DTW

For evaluating the Tesseract-DTW approach 6 photographs of different receipts were used. 5 of them were used as training samples and 1 as a test sample. This is repeated 6 times so each of the receipts was used once as a test sample.

First each character of the training receipts is labeled. Then feature extraction is done by using Tesseract. After the feature extraction each character of the learning receipts is described with 4 feature vectors (4). *X* and *Z* are with variable size (5) and *Y* and *W* are with constant size (6).

$$C_1 = (X^1, Y^1, Z^1, W^1) \qquad (4)$$

$$X^1 = (x_1, x_2, ..., x_m), Z^1 = (z_1, z_2, ..., z_j) \qquad (5)$$

$$Y^1 = (y_1, y_2, y_3), W^1 = (w_1, w_2, w_3) \qquad (6)$$

In the classification phase KNN classifier was used. For calculating the distance between two characters C_1 and C_2 combination of DTW and Euclidean distance measurement is used. DTW is used for calculating the distance between the vectors with varying size (7) and Euclidean distance is used for calculating the distance between the vectors with the no varying size (*8*). After DTW and Euclidean distance is calculated between the corresponding vectors of the two characters the final distance between the two characters is calculated with Euclidean distance based on the four distances (*d1, d2, d3, d4*) calculated in the previous step (9). The character with the smallest distance to the test character is chosen as the output of the classifier.

$$d1 = DTW_{dist}\ (X^1, X^2),\ d3 = \mathrm{DTW}_{\mathrm{dist}}\ (Z^1, Z^2) \qquad (7)$$

$$d2 = Euclid\ (Y1, Y2),\ d4 = Euclid\ (W1, W2) \qquad (8)$$

$$Distance\ (C_1, C_2) = Euclid\ (d1, d2, d3, d4) \qquad (9)$$

III. Experimental results

Figure 4 shows an accuracy comparison for the two approaches used for character recognition, Tesseract and Tesseract-DTW. The comparison is performed using the number of correctly recognized characters from 6 photographed receipts. One can note that only for the third photograph, the Tesseract-DTW is better than the original Tesseract. In all other cases the original Tesseract approach is better. On average, the Tesseract is better for 6 percentage points. Also compared by time of execution Tesseract was better that Tesseract-DTW, which was in a way expected given the complexity of the DTW and the usage of the so called "lazy" (instance-based) classifier – KNN.

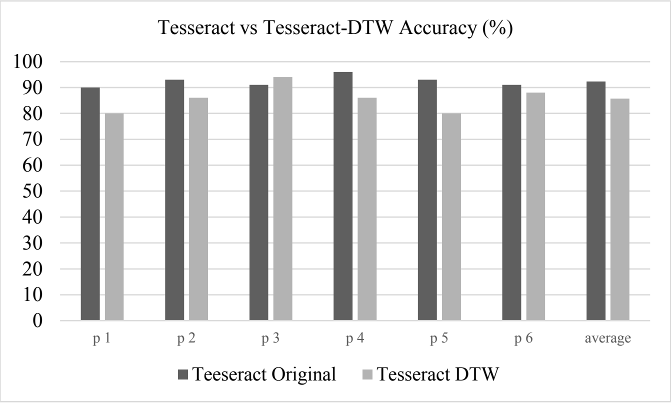


Figure 4: Accuracy for correctly recognized characters by using Tesseract and Tesseract-DTW.

IV. Discussion and Concusion

The paper presented an approach of OCR for receipts printed in Macedonian language. The main OCR engine that was used is Tesseract. In the process of character recognition two approaches were tested. In the first approach the original Tesseract was tested. Tests showed that Tesseract is most accurate when the training consists of random sampled words from different receipts and is written with similar font and size as the characters that we are trying to recognize. In the second approach modified version of Tesseract was used (Tesseract-DTW). In this approach the feature extraction was again performed by the Tesseract, however the final classification was done with KNN classifier using DTW as distance metrics. Tests showed that the first approach by using original Tesseract engine outperformed the second approach by 6 percentage points. Further analysis showed that the accuracy is higher for the numbers and words which are prescribed for each receipt, such as the date and the time of the purchase and the total amount of money spent. This was in a way expected because the classifier has more examples to train on, i.e. they are present in each receipt and the numbers are limited only to 10 characters. On the other hand, the names of the products are more difficult to recognize mainly because there are names that are not from Macedonian language. In general, the more data is used for training, the better the model should be. In future we plan to collect much more data samples by providing a free smartphone application.

To the best of our knowledge, this is the first attempt to apply OCR on receipts printed in Macedonian language using the Cyrillic script, and moreover the first attempt to modify the original Tesseract by applying KNN algorithm using DTW distance metrics. Even though, the modified version of the Tesseract achieved slightly worse results, it gives promising results and we plan to further improve it in the future work. We are also considering an approach that will combine the both methods, e.g. by using meta-learning, and eventually improve the recognition accuracy.

Acknowledgment

The authors would like to thank the developers of the Tesseract OCR framework and making it freely available to the research and developers community.

References

[1] G. Tauschek, “Reading machine” U.S. Patent 2026329, Dec. 1935

[2] S. Mori, C.Y. Suen, K. Yamamoto“Historical Review of OCR research and development”, Proceeding of the IEEE (Volume:80, Issue 7), Jul 1992

[3] Google’s Goggle application.
https://play.google.com/store/apps/details?id=com.google.android.apps.unveil

[4] Google’ API for OCR.
https://developers.google.com/google-apps/documents-list/#uploading_documents_using_optical_character_recognition_ocr

[5] Thanslang application.
https://play.google.com/store/apps/details?id=icactive.app.translang

[6] Tesseract-ocr. Mar-2012. URL: http://code.google.com/p/tesseract-ocr/.

[7] Ray Smith. “An overview of the Tesseract OCR engine". In: Document Analysis and Recognition, 2007. ICDAR (2007).

[8] M. Müller, “Information Retrival for music and motion”, 2007, XVI, 318 p. 136 illus. 39.

[9] D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.

[10] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 26, no. 1, pp. 43–49, 1978.

[11] Y. Qiao, X. Wang, C. Xu, “Learning Mahalanobis Distance for DTW based Online SignatureVerification”, Information and Automation (ICIA), 2011 IEEE International Conference, June 2011

[12] A. Piyush Shanker, A.N. Rajagopalan, “Off-line signature verification using DTW”, Journal Pattern Recognition Letters Volume 28 Issue 12, September 2007

Testing web applications

Aleksandra Angelovska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia

Anastas Misev
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia

***Abstract* — We are living in a world in which everything is based on Internet. Because information shared online is available at any time and any place, many businesses right here see the opportunity of cheap and easy advertising and promotion of what they have to offer. The applications that are made today are increasingly complex, have a myriad of functions and therefore require a more sophisticated way of testing them. Because the future of many businesses relies on web application, the company cannot afford making public application with errors which would affect thousands of customers and loss of significant funds. The different nature and complex character of the functionalities that one web application can offer, because today's web applications are complete software solution, leads to make more difficult the process of testing and quality assurance. This paper is an overview of what so far has been used as a method for testing web applications, is a description of advanced testing methods which are less represented, and aims to give direction on how tester should correctly choose a method that for shorter period of time will pass larger amount of test scenarios, and yet the application will be quality tested, and a lot of errors will be found to be fixed before launching the application. Only the correct choice and combination of test methods improves the tester and finally delivers quality application.**

Keywords—web application; testing; quality assurance; testing methods

I. INTRODUCTION

The only way to ensure that some web application will function properly is to test it. Since there is rapidly growing demand for web based application and companies want to deliver the solution on time, testing of the product is often considered as process that can cost a lot of money that company cannot afford, so, sometimes they launch the product that is untested, and with lots of bugs that are not found by the developers in the developing process. Due to that approach, many companies instead of gaining money for delivering product on time, with all functionalities implemented, they are losing money because they deliver a product that does not work properly.

Clients do not want web applications that are delivered to them to be dysfunctional, aren't ready on time, do not meet requirements defined in the BRD, have bugs or are too complicated to use. So, companies need a QA team to perform several types of tests to ensure the clients that product that will be delivered to them will be fully functional and they will get what they paid for.

Although much is known about the need of testing the application before delivering, and there are a lot of techniques that are offered, so little is invested in that process, and yet, the blame is all on the QA when production bugs are returned.

This paper will explain the current techniques that are used for testing web based applications, and will explain some techniques that are less used so it is up on the QA to choose the perfect set of methods to be efficient and in limited period of time to test entire application correctly.

II. BACKGROUND

Web application is software that is accessible through web browsers. It is a complex type of software available online, that has complex graphical user interface and back-end software components. Web applications are client-server type of applications that can be written in several languages, often communicate with databases and make the process of testing them more difficult and complex.

"The aim of Web application testing consists of executing the application using combinations of input and state to reveal failures". [1]

Therefore, there are a lot of methods that can be used to find more bugs before the final release of the product.

A. *Unit testing*

The Unit test is a small piece of code written by developers, mostly before starting with code writing, to ensure that code that will be written later will pass. If the unit test fails, then the developer will be able to rewrite the code, to make the test to pass, and to eliminate most of the bags that later will be reported by the QA. It is important to write good unit test that can be repeatable, not after small change to be forced to rewrite the test in order to pass.

Unit tests are only small piece of the testing process in bug elimination. Tests that pass are not a guarantee that that functionality will be without bugs.

B. *Integration testing*

Integration testing is a testing process that tests application modules and assures how those modules interact with interfaces. It is a functional type of testing and according to [5, 6] covers:

- Calls of different software components, while interacting to each other
- Data and information sharing between the modules in proper manners

- Compatibility, which ensures one module that does not effect on the performance and functionality of the other modules.
- Nonfunctional issues

For making good and quality integration tests, the knowledge of both the structure and the behavior of the Web application will have to be considered, so is more likely to be performed by the developers than the QA team.

It is important to be done, but many companies do not perform integration tests if the web application that is developed is extremely critical.

C. *Manual testing*

Manual testing is a widely spread method for testing of web applications. It is a method for testing with execution of positive and negative scenarios of some functionality in order to find a gap where that functionality will not work. Because it is performed by the QA team, it spends a lot of resources and is expensive method of testing.

Manual testing is an easy way of testing that does not include code, scripts, algorithms, but is highly error prone, since the missed steps or incorrect understanding of some functionality can leads to false results.

Manual testing also includes cross browser testing and testing the application on different devices like different operating systems, desktop computers, laptops and mobile devices.

The problem of manual testing, beside the spending lot of time with test case running is accurate defect reporting, because sometimes test case can have false positive result of the expected result.

But even though applications became most complex and powerful, still manual testing has remained as a method that is most common for validating the functionality and assuring of its quality. There always be parts of the application that should be tested manually, and where quick manual testing (like smoke test) will perform enough data for some functionality.

D. *Load and performance testing*

Load tests are type of tests that are performed in order to measure the response time of the web application, and how quick customers can access that application.

Because today many business rely completely on web application, that application should be available nearly 100 percent of the time, to avoid losing money.

Web application performance testing is a priority testing before going live. QA team creates load and performance tests mostly by creating virtual users and helps them measure the behavior of the application.

High web application traffic, complex and numerous functionalities and a lot of changes due to development process are good indicators that load and performance tests should be created and executed.

There are a lot of load and performance test tools that can help the QA team with their measures, and can deliver a correct data for analysis to the developers, so they can re-factor the code to improve the performance of the application before the release.

E. *Security testing*

"Vulnerabilities in web applications are now the largest source of enterprise security attacks". [9] The most common ones: "cross-site scripting," "SQL injection," and "buffer overflow" can make a real trouble to the company especially to the sensitive data of the application.

The rapid growth of the web applications through years has created systems that are easy target that hard can be secured. To deliver application that user can rely on, and can be confident to use, without fear of security issues, application should be tested and secured from several types of security vulnerabilities according to [9]: stealing user account identities, illegal access to applications, hijacks, expose of sensitive data, interfere with application usage, etc.

Since traditional manual testing cannot detect those vulnerabilities, DBAs and developers should help on detecting and preventing those vulnerabilities in order to deliver secure web product.

There are many tools on the market that will help detection on those issues, and this kind of test should be not excluded from project development process, and should be performed before launch of the web application, since the Internet data is easy target of security attacks.

F. *Regression testing*

Regression testing is a process of testing performed by QA team in order to ensure that changes made during the development lifecycle of the application have not affected the features that were not concerned by the change. This type of testing is performed when all development tasks are closed, and QA has tested the application's functionalities once.

But because takes a lot of time to run the test cases again, especially when large web application is tested, sometimes some of the test cases are excluded from the regression process in order to finish the work on time. The selection criteria of which test cases will be performed in the regression testing process depends on many factors, but mostly from the functionalities affected by the changes and the bugs that are reactivated several times, when something new is added.

Many companies does not cover regression testing at all since it costs a lot of money, resources, effort and time, but importance of this type of testing is extremely large, because a lot of bugs can be found that can be fixed before release.

III. ADVANCED TECHNUIQES FOT TESTING WEB APPLICATIONS

Web applications are more and more complex, the need for them is growing every day; QA team must find more appropriate and more sophisticated way of testing those web applications. Since many companies develop the product using agile methodologies which exclude the testing process, QA team must find more suitable way of testing the web applications in a limited period of time.

Automation and mutation testing are two types of testing that in short period of time can cover a lot of application's functionalities.

A. *Automation testing*

Automation testing is a process of writing coded UI tests that can be executed repetitively in order to test functionalities that are not changing, and when manual testing will take a lot of time.

Since delivering the product after the deadline can lose money and customers, make many companies to adopt automation to reduce time and testing budgets. But better to be late than to launch buggy application, so it's not always great solution to choose automation instead of other testing techniques.

Automation is more than saving the steps in the browser and executing the tests all over again. It should be done carefully so the tests can be valid, and more bugs to be found.

Automation should not be performed when the application is unstable, when the testers are not familiar with the application or are inexperienced.

But even if automation testing is process of testing that reduces the time and cost, if there are not enough people to perform the testing automation won't reduce the errors.

So, automation testing is a good type of testing that reduces a lot of time that many companies cannot afford due to application development lifecycle, should not be performed as one and only testing method, but if the tests are good, can find a lot of bugs.

B. *Mutation testing*

Mutation testing is a testing method that is not much used nowadays in the companies as method that assures the quality of the application.

Mutation testing uses a piece of program, modifies that piece, and helps the QA by entering input data to detect whether that program behave differently from the original or not. The modified versions are called mutants and if modified version acts differently than the original than that test kills the mutant. [14] "The idea of using mutants to measure test suite adequacy showed empirical support for one of the basic premises of mutation testing, that a test data set that detects simple faults (such as those introduced by mutation) will detect complex faults, i.e., the combination of several simple faults." [13]

According to [13] there are four classes of mutation operators:

- Replace an integer constant C by 0, 1, -1, ((C)+1), or ((C)-1)
- Replace an arithmetic, relational, logical, bitwise logical, increment/decrement, or arithmetic-assignment operator by another operator from the same class
- Negate the decision in an *if* or *while* statement
- Delete a statement

By making those changes in the program, tester can isolate dysfunctional piece of code.

Only experienced testers should perform mutant testing, and when is performed carefully and qualitative, can provide a good indication of the fault detection ability of a test suite.

IV. HOW TO CHOOSE APPROPRIATE TEST METHODS

Each testing technique is good and bad on its own way in the process of finding certain kind of bugs and issues that provides dysfunctional functionalities. Choosing the right and appropriate testing method is up to the testers and does not always relies completely on them.

According to [15], there is certain kind of schema that should be followed in selecting appropriate testing technique:

TABLE I. CHOSING APPROPIATE TESTING TECHNIQUE

Element	***Attribute***	***Description***
Technique	Comprehensibility	Whether or not the technique is easy to understand
	Maturity level	How experimental and/or how well validated the technique is
	Cost of application	How much effort is needed to apply the technique
	Inputs	Inputs required to apply the technique
	Adequacy criterion	Test case generation and stopping rule of the technique
	Test data cost	Cost of identifying the test data
	Dependencies	Relationships of one technique to another
	Repeatability	Whether two people generate the same test cases
	Sources of information	Where to find information about the technique
Results	Coverage	Coverage provided by the set of test cases
	Effectiveness	Capability of the set of cases to detect defects
	Type of defects	Type of defects the technique helps to discover
	Number of generated cases	Number of test cases generated per software size unit
Object	Phase	Stage of development at which the test is to be run
	Element	Elements of the system on which the test acts
	Aspect	Functionality of the system to be tested
	Software type	Type of software that can be tested using

Element	*Attribute*	*Description*
		the technique
	Software architecture	Development paradigm to which the technique is linked
	Programming language	Programming language with which the technique can be used
	Development method	Development method or life cycle to which the technique is linked
	Size	What size the software should be to be able to use the technique
Tools	Identifier	Name of the tool and the manufacturer
	Automation	Part of the technique automated by the tool
	Cost	Cost of tool purchase and maintenance
	Environment	Platform (sw and hw) and programming language with which the tool operates
	Support	Support provided by the tool manufacturer
Agents	Experience	Experience required to use the technique
	Knowledge	Knowledge required to be able to apply the technique
Project	Reference projects	Earlier projects in which the technique has been used
	Tools used	Tools used with the technique in earlier projects
	Personnel	Personnel who used the technique on earlier projects
Satisfaction	Opinion	General opinion about the technique after having used it
	Benefits	Benefits of using the technique
	Problems	Problems with using the technique

V. CONCLUSION

This paper has introduced a lot of techniques that can be used by the quality assurance team for performing the tests in order to deliver a product that accomplish the requirements that are ordered by the clients.

It is given an overview and traps are observed that can occur when using a specific method especially in short period of time, when the QA team should work under pressure in order to cover with testing entire application, and yet to find all the possible bugs.

By following the best practices and the most appropriate testing methods, can be reduced the risk of failures, can be delivered product that accomplish the requirements, can be increased the quality of the application, and the QA team can be fast, precise, and more organized.

REFERENCES

[1] Giuseppe A. Di Lucca, Anna Rita Fasolino, “Testing Web-based applications: The state of the art and future trends”, 22 August 2006

[2] Roy Osherove, “The Art of Unit testing with examples in .NET”, second edition, 2013

[3] Tim Mackinnon, Steve Freeman, Philip Craig, “Endo-Testing: Unit Testing with Mock Objects”, 2000

[4] Umar Farooq, Usman Azmat, “Testing Challenges in Web-based Applications with respect to Interoperability and Integration”, January 2009

[5] Imran Akhtar Khan, Roopa Singh, “Quality Assurance And Integration Testingaspects In Web Based Applications”, June 2012

[6] HP Business White Paper, “How to increase the efficiency of manual testing”, November 2010

[7] Daniel A. Menasce, “Load testing web sites”, August 2002

[8] HP Business White Paper, “An introduction to load testing for web applications”, July 2010

[9] QUALYS, “Web Application Security - How to Minimize the Risk of Attacks”, February 2011

[10] Gregory M. Kapfhammer , “Regression Testing”

[11] Christopher Jerry Mallery, “On the feasibillity of using FSM approaches to test large web applications”, May 2005

[12] Linda G. Hayes, “The Automated Testing Handbook”

[13] J.H. Andrews, L.C. Briand, Y. Labiche , “Is Mutation an Appropriate Tool for Testing Experiments?”

[14] Upsorn Praphamontripong, Jeff Offutt, “Applying Mutation Testing to Web Applications”

[15] Sira Vegas, Victor Basili, “A Characterization Schema for Software Testing Techniques”

Using the Spring Framework with Java Enterprise Edition when Creating MVC Platform

Marjan Tanevski, Adrijan Bozhinovski, Eva Blazhevska, Biljana Stojchevska
School of Computer Science and Information Technology
University American College Skopje, Macedonia
marjantanevski@outlook.com, {bozinovski, blazevska, stojcevska}@uacs.edu.mk

***Abstract* — With the rapid development of network technologies and high competition for the development of web applications, there is an increasing demand for such applications and also a rise in their complexity in terms of implementation. This paper describes the ideas of the Spring framework commonly used to develop business network applications more accurately and efficiently. The Spring framework is a complex framework that consists of other schemes which are integrated together, and some of which will be described and discussed in this paper, such as Spring MVC, IoC, and AOP. It will be described how the MVC application framework of Spring is implemented in the J2EE platform and the MVC design pattern will be taken into consideration, which is important to separate the dependencies between components. The initial phase of the development of web applications will be described, first without the MVC model, and its development will be followed until the Spring MVC. IoC (Inversion of Control) is one of the many features of Spring that will be considered as an essential part of the Spring and programming technique in object - oriented programming with AOP (Aspect-Oriented Programming) as a programming paradigm applied to the basics of the Spring, which forms the bases for further development of the aspect-oriented software.**

Keywords: Spring framework, J2EE, MVC, IoC, AOP

I. Introduction

With the introduction of information technology and communication media, many of the companies already use frameworks to make the development of their applications easier. The business today demands web applications to advertise its company so it is very important to take care of the architecture used in the development of the application. Bawiskar et al; [1] suggest that the Spring application framework is a kind of framework that helps to adapt Java applications effectively. Johnson et al;[2] and Zhao et al;[3] explain that even though JavaEE is widely used, it has some limitations such as code reusability, which is a heavy burden in the development of the application. Therefore, this platform requires frameworks such as Spring. Bawiskar et al; [4] indicate that having the Spring framework together with JavaEE facilitates the development of software (Java EE - Java Platform, Enterprise Edition or Java EE is Oracle's enterprise Java computing platform [5]). Ju and Bo [6] describe Spring as a free and open source framework that offers a number of features for developers. The first version of this framework was created by Rod Johnson [7].

Spring is a layered architecture, so whenever an E-commerce system is developed, a clear overview of the separation of the layers must be maintained [4]. Such layered architecture allows the users to select which of its components to use and which not. Johnson et al; [8]suggest that the Spring architecture consists of seven modules which can be shown as follows:

- Core container
    - Beans, Core, Context, SpEL,
- Data Access/Integration
    - JDBC, ORM, OXM, JMS – layers above Transactions,
- Web
    - WebSocket, Servlet, Web &Portlet,
- AOP (Aspect-oriented programming, Aspects, message exchange) and instrumentation, and
- Test

The modules are shown on Figure 1.

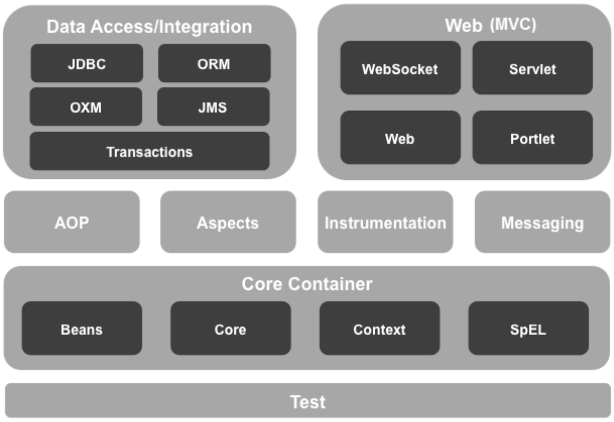


Figure1:Springlayered architecture [9]

Spring is based on two main components: inversion of control (IoC) and aspect-oriented programming. Gupta and Govil [10], however, indicate that besides these two main components Spring has its own MVC model. In this paper the inversion of control and aspect-oriented programming will be describedbriefly, while Spring MVC will be presented by past research that has been done on this scheme.

II. ASPECT-ORIENTED PROGRAMMING

Ju and Bo [6] suggest that AOP (Aspect Oriented Programming) is a technique which separates the crosscutting concerns in one system. Aspects can be transactions, logs and security. In Spring, aspects are merged together with the help of the Spring configuration file, which is an .xml file, which makes the framework more modular. Johnson et al;[8]in the documentation for Spring explain that AOP in Spring is used to:

- provide declarative business services (e.g. declarative transaction management) and
- let users to implement custom aspects

III. INVERSION OF CONTROL

According to Graudins and Zaitseva [11], the Java 2 Enterprise Edition (J2EE) platform, which provides great opportunities for distribute systems development, is used for the most enterprise applications and Enterprise JavaBeans (EJB) technology is the heart of it; bean or JavaBean is a class that encapsulates many objects into a single object (the bean) [12]. Usually the architecture of a J2EE application contains several separate layers (as shown of Fig. 2).

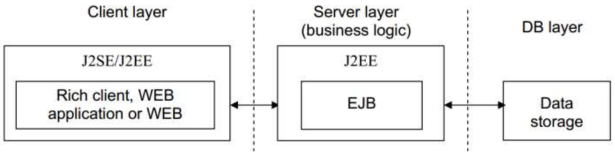

Client layer
Server layer (business logic)
DB layer
J2SE/J2EE
Rich client, WEB application or WEB
J2EE
EJB
Data storage

Figure 2: Classic 3-tier architecture[11]

According to Graudins and Zaitseva [11], the Server layer typically contains server components with application business logic, which are managed by an EJB container [13].The EJB container is a part of the application server (typically the EJB container and the application server cannot be separated and are produced by the same vendor). It provides a server component lifecycle, as well as transaction and security management services.

The EJB specifications are intended to provide a standard way of implementing back-end business code which is found in business applications (against front-end interface code)[13]. Both types of code face similar problems and solutions to these problems are often repeatedly implemented by developers. Enterprise JavaBeans are intended to handle common concerns such as persistence, transactional integrity and safety in a standard way, leaving programmers free to concentrate on a given problem. According to the documentation by Oracle Corporation[14],one of the main goals of the EJB architecture is to facilitate the writing of distributed object-oriented business applications in the Java programming language. Lambert [15], however, indicates that the EJB versions 1.0 - 2.1 were too complicated and did not achieve this goal, but the purpose of the EJB 3.0 release is to improve the EJB architecture by reducing its complexity.

Because earlier version of EJB were too complicated, new technologies appeared to manage the business components (such as Spring), which can be applied with EJB and instead of EJB. Walls and Breidenbach [16] point that the main goal of the Spring Framework producers was to create a simple alternative to EJB. Usually, objects obtain references to required objects by themselves. Graudins and Zaitseva [11] point that inversion of control allows injecting all dependencies into beans during creation time by an external manager. A bean is only required to define a required property in the code and its mutator (i.e. set()) method. A primary source for dependency injection is the xml configuration file like the one shown on code 1:

```
<beans>
<bean id="customerService" class="com.article.CustomerServiceImpl"/>
<bean id="productService" class="com.article.ProductServiceImpl">
<property name="customer" ref="customerService"/>
</bean>
</beans>
```

Code 1:Dependency injection

IV. SPRING MVC

According to Gupta and Govil [17], the development of the MVC pattern has three steps: without MVC, the MVC model 1 and MVC model 2.

A. *No MVC*

In the initial phase of web applications development, the pages were used to be designed in HTML, which is plain text only. This was the first markup language which was able to work on the Internet. And today it still works as a building block for the all Internet based programming languages. The user had to interact with the static pages. The information written on the pages had to be changed manually. As the time passed, the demand arose for a language that could interact with the user and pages could be changed per the requirement of the user.

B. *MVC model 1*

The first major change in the architecture came with the introduction of the MVC Model 1 Architecture. This architecture was completely based on the page-centric approach. In this model, a Java Server Page was used to control the Presentation, Business Logic and flow of the program. In this model, the concept of the Business Logic was introduced, which was hard-coded in the form of Java Beans and scriptlets. The entire code was written within a JSP page. Figure 3 presents a case where the flow of a JSP application based on the data received from the input needs to be transferred.

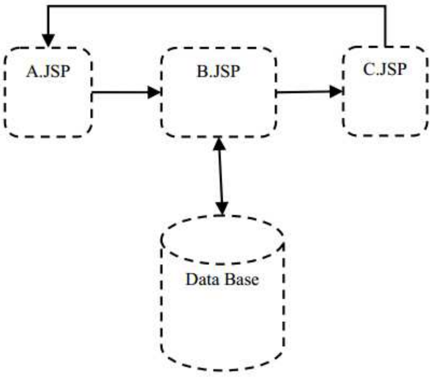


Figure 3: Page Navigation in the MVC -1 architecture [17]

C. *MVC model 2*

The model 1 architecture was able to solve some of the problems of the web and Internet programming but still there were lot of things missing from it. It was centered on the navigation of the JSP pages so there was the challenge of further development in the architecture point of view. During this process, the next step was the development of the Model 2 architecture. This problem was solved using the Servlet and JSP together.

The Servlet handled the initial request and partially processed the data. It set up the beans then forwarded the result to one of the JSP pages. The Servlet decided the page to be displayed from the list of available pages.

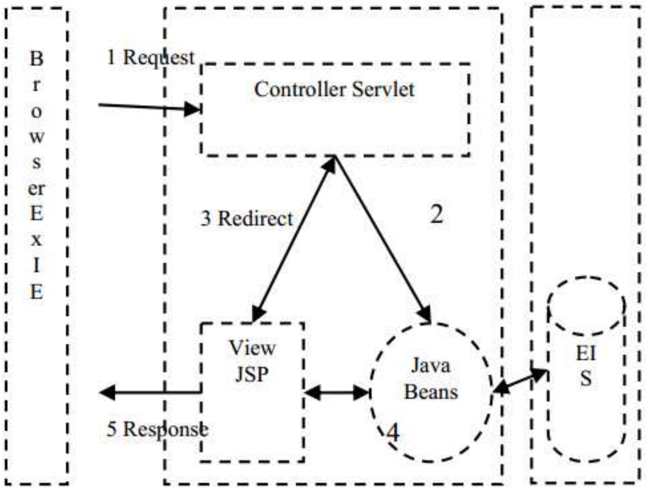


Figure 4: MVC 2 architecture[17]

In the Model 2 architecture, JSP Pages were used for presentation purposes only were the Business logic was removed from the web page. This made the pages easier to display and maintain.

In this model, all control and application logic was handled by the Servlet. The Servlet was written in the Java programming language, so it was also easier to handle the programming part of the Servlet. In this scenario, the Servlet became powerful enough for the complete application and emerged as the central point for the application.

In the model 2 architecture, the Servlet became the gatekeeper for the all common tasks. It provided the common services like authentication, authorization, error control and flow of the application. This architecture solved most of the problems. But still there were many new issues that emerged during the application of this architecture.

D. *Spring MVC main components*

In the Spring architecture, the principles of MVC are also followed. It has been designed for most of the desktop and Internet based applications. According to Gupta and Govil [10], MVC contains three main components:

Controller: Handles navigation logic and interacts with the Service tier for business logic.

Model: Represents the contract between the Controller and the View that contains the data needed to render the View populated by the Controller.

View: Renders the response to the request by pulling appropriate data from the model.

Core components in the Spring MVC are as follows:

DispatcherServlet: This is the Spring's front controller implementation. When web.xml receives the request, it transfers it to the DispatcherServlet. The Controller is the first to interact with the requests. It is also known as the implementation of the Servlet. It controls and navigates the complete flow of the application.

Controller: It represents the user created components for handling requests and encapsulates their navigation logic within them. It also delegates the services for the service object.

View: It is responsible for rendering the output. Different views can be selected for the different types of output based on the results and the viewing device, such as various communication devices.

ModelAndView: It is the core part of the Spring framework. It implements the business logic of the application. It is created by the controller and associates the view to the request. It stores the business logic and Model data and is called by a controller, which controls its execution. After the execution, it returns the data and name of the view.

ViewResolver: How the output is to be displayed depends on the results received from ModelAndView, whereas the ViewResolver is used to map logical view names to actual view implementations. This part is used to identify and implement the output medium type and the appropriate display to suit it.

HandlerMapping. Represents a strategic interface used by the Dispatcher Servlet for mapping incoming requests to individual Controllers. It identifies the request and calls the respective handler to provide the services. The Handler will in turn call the appropriate controller.

E. The Spring MVC architecture

The Spring MVC is configurable with multiple view technologies such as Java Server Pages, Velocity, Tiles, iText, Tapestry, Wicket, SiteMesh etc [10]. The Spring MVC separates the roles of the controller, model object, Dispatcher Servlet and the handler object. Such clear separation of objects and controllers makes them more easily customizable. Figure 5 shows the flow of execution.

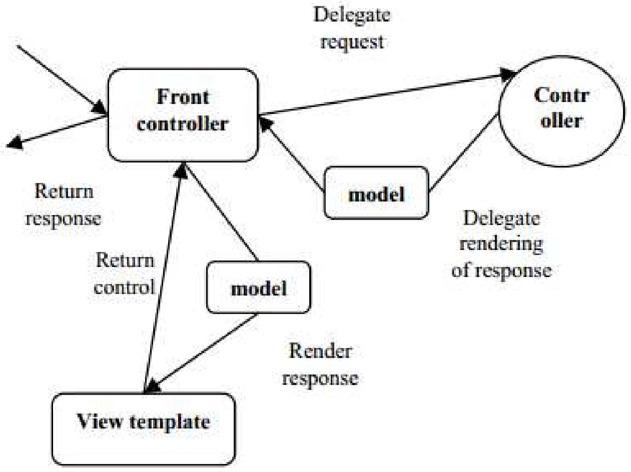


Figure 5: Sequential diagram of the execution flow of the Spring MVC[10]

Figure 5 shows the sequential diagram of the Spring model, where Figure 6 shows the sequential flow. In it, the DispatcherServlet is the entry point for the application. As soon as the DispatcherServlet gets the request for the services, it will decide the handler. All handlers are mapped with the Servlet. The role of the handler is to call the appropriate controller and to pass the request parameters to it.

Because the controller contains the business logic and a ModelAndView is associated to it, it will return the ModelAndView to the DispatcherServlet during execution time. The ModelAndView contains all the required data and the name of the view.

Once the DispatcherServlet gets the ModelAndView from the controller, it calls the appropriate view resolver. The view resolver will identify the name of the view using the data which need to be presented. In the end, the data will be displayed to the user according to the requested user format.

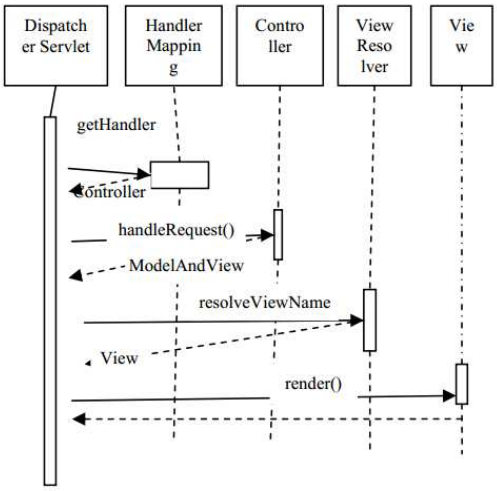


Figure 6: Sequential flow of the application in Spring MVC [17]

V. ARCHITECTURAL BENEFITS OF THE SPRING FRAMEWORK

There are many architectural benefits of the Spring framework which can contribute to a project [10]. Following is a list of some of them:

- Spring effectively organizes the middle tier objects. The configuration management services can be used in any architectural layer and in any runtime environment;
- The Spring Web MVC Framework is a robust, flexible, and well-designed framework for rapidly developing web applications using the MVC design pattern;
- Clear separation of roles: the Spring MVC nicely separates the roles taken by the various components that make up this web framework. All components like controllers, command objects, and valuators have distinct roles;
- Adaptable controllers: If the application does not require HTML forms, a simpler version of a Spring controller can

be written, that does need all the extra components required to form controllers. Spring provides several types of controllers, each serving a different purpose;

- Spring eliminates the need to use a variety of custom types of file formats, by handling configuration in a consistent way throughout applications and projects;
- Spring provides sound programming practice by reducing the programming cost to interfaces, rather than classes;
- Applications built with it depend on a few of its APIs. Most of the business objects in Spring applications are not dependent on Spring;
- Applications built using Spring are very easy to unit test;
- Spring can make use of EJB as an implementation choice, instead of as a determinant of the application architecture;
- The user can choose to implement business interfaces as POJOs (Plain Old Java Objects) or local EJBs without affecting the invocation code;
- Spring provides an alternative to EJB that is more appropriate for many applications. It can use AOP to deliver declarative transaction management without using an EJB container;
- The Spring Framework can be effectively used with other frameworks such as Struts, Hibernate etc.;
- Due to its Inversion of Control feature, the amount of time needed for testing the code is significantly lowered;
- Because Spring is a layered architecture, the users can select which of its components can be used;
- The Spring Web MVC provides controllers to ease dealing with various HTTP requests from the user;
- The Spring Framework can work efficiently with J2EE for developing applications in an effective manner;

VI. Conclusion

The Spring framework is a powerful framework for building enterprise-wide Java applications. Using it, the application can rely on consistency, performance and reliability. The IoC and AOP features promote good programming habits, while Spring finds an easy way to use the EJB application. It can also be easily integrated with some other frameworks such as the Struts and Hibernate frameworks for developing efficient enterprise-wide Java applications, thereby reducing the coupling with clear separation of layers, making them easier to understand. SinceSpringis an open source environment, it is recommended that the developers make use of this technology for the development of large size web applications.

VII. References

[1] Bawiskar, P. Sawant, V. Kankate, Dr. B. B. Meshram K. Elissa (2012) Integration of Struts, Spring and Hibernate for an University Management System, International Journal of Emerging Technology and Advanced Engineering, vol. 2, n. 6, pp. 2250-2459.

[2] M. R. Johnson, J. Hoeller, A. Arendsen, T. Risberg, C. Sampaleanu (2005) Professional Java Development with the Spring Framework, John Wiley & Sons Publications.

[3] Zhao, M. Jiang, Z. He (2010) The Design of E- Commerce System Architecture Based on Struts2, Spring and Hibernate, IEEE, pp. 3251 – 3254.

[4] Bawiskar, P. Sawant, V. Kankate, Dr. B.B. Meshram (2012), Spring Framework: A Companion to JavaEE , IJCEM International Journal of Computational Engineering & Management, vol. 15 n. 3, pp. 2230-7893.

[5] Ian Evans (2012), “Differences between Java EE and Java SE -Your First Cup: An Introduction to the Java EE Platform”, http://docs.oracle.com/ [last accessed on 30 March 2014].

[6] K. Ju, J. Bo (2007) Applying IoC and AOP to the Architecture of Reflective Middleware, IFIP International Conference on Network and Parallel Computing - Workshop, pp. 903-908.

[7] C. Vijayakumar (2011), Analysis of the Frame work Standard and the Respective Technologies, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol.1, No.1, October 2011, pp. 42-48

[8] M. R. Johnson, J. Hoeller, A. Arendsen (2014) Spring. Java/J2EE Application Framework, http://www.springframework.org/documentation [last accessed on 10 February 2014].

[9] M. R. Johnson, J. Hoeller, K. Donald, C. Sampaleanu, R. Harrop, T. Risberg, A. Arendsen, D. Davison, D. Kopylenko, M. Pollack, T. Templier, E. Vervaet, P. Tung, B. Hale, A. Colyer, J. Lewis, C. Leau, M. Fisher, S. Brannen, R. Laddad, A. Poutsma, C. Beams, T. Abedrabbo, A. Clement, D. Syer, O. Gierke, R. Stoyanchev, P. Webb, R. Winch, B. Clozel (2013) Overview of the Spring Framework, http://docs.spring.io/spring/docs/4.0.2.BUILD-SNAPSHOT/spring-framework-reference/htmlsingle/#overview-modules [last accessed on 10 February 2014] .

[10] P. Gupta, Prof. M.C. Govil (2010) Spring Web MVC Framework for rapid open source J2EE application development: a case study, International Journal of Engineering Science and Technology, vol. 2, n. 6, pp.1684-1689.

[11] J. Graudins, L. Zaitseva (2006) Comparative Analysis of EJB3 and Spring Framework, International Conference on Computer Systems and Technologies - CompSysTech, pp. 18: 1-4.

[12] P. Coffee, (1997) How to program JavaBeans. Emeryville Publications.

[13] Y. Song, N. Wei (2013) “Introduction to J2EE”, http://www.ics.uci.edu/~cs237/presentation2013/J2EE.pptx [last accessed on 19 February 2014]

[14] Oracle Corporation (1999) Enterprise JavaBeans, http://docs.oracle.com/cd/F49540_01/DOC/java.815/a64683/ejb.htm [last accessed on 21 February 2014].

[15] R. Lambert (2005) An Introduction to the Spring Framework, http://www.developerdownload.com/113602/ [last accessed on 19 February 2014].

[16] C. Walls, R. Breidenbach (2005) *Spring in Action*, Manning Publications

[17] P. Gupta, Prof. M.C. Govil (2010) MVC Design Pattern for the multi framework distributed applications using XML, spring and struts framework, International Journal on Computer Science and Engineering, vol. 2, n. 4, pp. 1047-1051.

Analysis of the usage of social addons in viral marketing in Republic of Macedonia

Tomche Delev*, Dejan Gjorgjevikj†
Faculty of Computer Science and Engineering
Skopje, R. Macedonia
* tdelev@finki.ukim.mk
† dejan.gjorgjevikj@finki.ukim.mk

Igorcho Donchovski
doncovski_igor@yahoo.com

***Abstract*—In this paper we analyze the contribution of using social addons in viral expansion of news on the Internet. We investigate the hypothesis that one web page can become more popular if it's more social, by enriching the interaction with the users by integrating social addons. The research is focused on the Internet users in Republic of Macedonia and the goal is to predict the future usage of social addons and their effect on viral marketing. The results in this paper concluded from the answers of the two questionnaires, are giving answers to questions such as: the time spend browsing social sites, location and purpose of visiting social sites, and if the time spent od social sites is corelated with using social addons. At last it answers the quesiton of faster or viral spreading of news and stories , if they are shared using social addons.**

I. Introduction

In todays era of social networks, the first instinct of the users when they see or read something interesting on a web page, is sharing. The social web itself is built on the idea of sharing things between users. The goal is to enable and embrace users to share their life using the social networks. They share their emotions, important life moments, relationship status by posting text, pictures or videos.

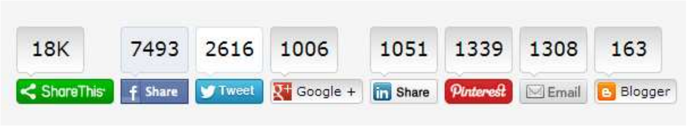


Fig. 1. Social plugins.

Social addons (shown of figure 1) are just another step to simple and ubiquitous way of sharing things from web pages. They are simple hyperlinks with icons that allow users to share, recommend, rank or just comment any article they read on a web page. The owners of the web sites are using the social addons to allow easy sharing of their content. The expectation of the owners is generation of more traffic, increased page rank and hence increased popularity of their site. Integrating social addons is performed by including snippet of HTML and JavaScript code in the source of the web page. Interaction between users and social addons is explained with these few steps:

1) users click on some addon link
2) if the user is not signed in the addon web site, then sign in form is shown
3) the form with the sharing content is automatically filled and the user is allowed to modify before posting.

For each piece of content that is shared, the social sites are providing information of the number of sharings of that unique content (usually identified by the URL). This information identifies the most popular stories or web pages.

The sharing of content with social addon has the possibility to create the Slashdot effect. The possible traffic generated to the web page can be described as: sharing brings page visits, visits bring conversions. Traffic volume measures give no indication of whether the audience referred to the site engages with it, so we need quality measures to show this. *Conversion rate* is the best known quality measure which shows what proportion of the visitors from different sources within a defined time period convert to specific marketing outcomes on the web, such as lead, sale or subscription. Example: 10% of visitors convert to an outcome such as logging in to their account, or asking for a quote for a product. Conversion rates can be expressed in two different ways - at visit level (visit or session conversion rate) or the unique level (visitor conversion rate).

Sharing from users, the visits and conversions can be used defining three coefficients usable in grading specific social networks and the effect of social network on business.

$$\frac{Conversions}{Social\ Actions} \tag{1}$$

1 How often social sharing leads to conversion.

$$\frac{Conversions}{Social\ Visits} \tag{2}$$

2 How often shared links on social networks leads to conversion.

$$\frac{Social\ Visits}{Social\ Actions} \tag{3}$$

3 How often sharing on social networks leads to increased site traffic.

II. Social networks and internet marketing

A. Social networks

There are many definitions for social networks [1]. According to one of them social network is sociable structure from members or organizations, called nodes, connected with one or more specific interrelations such as friendship, relatives, or any other common interest. Social networks gives the people opportunities to share information, give support to each other and in general are very important part in their lives [2]. Extending the definition of social networks, the social web site [3] or social media is a web site that creates virtual community for people to share their daily activities with family and friends, or to share their other interest. The thing that makes the social sites unique, is that they enable the users to articulate and make visible their social connections [4]. This can result in connections that aren't feasible, but this is not the point. Social sites by definition are enabling new type of communication, where the computer is the basic device for collaboration between groups [5]. Social sites are a cyber space, that enables the users to build their virtual profiles, to share text, photographies, video and to connect to other users of the site. On most of the large social network sites, the users are not joining to find or meet new people, but to connect to ones they already know and are part of their real life social network [6].

B. Internet marketing

Recently, many researchers try to define the term Internet marketing. The Internet marketing is an internet application, used to achieve the original goals of marketing. It can also be defined as usage of digital technologies to achieve marketing goals [7].

Marketing on social media is a term that describes the usage of social network sites, on-line communities, blogs or any other social media in marketing, sales, public relations or customer service.

C. Viral marketing

Creating a web site is only one part of the process of internet presence and successful advertising. The question of your popularity, or the amount of people that will notice your name or business is open question. Recently Google incorporated so called "Social search" that will search for activities in social network sites. The novelty aspect is the search in contents of social sites such as comments, likes, recommendations. This enables new type of viral marketing using social media. It is a marketing phenomena where a marketing messages is spread by sharing from social sites users. When the spreading of the news is fast like virus, it's called viral. The viral marketing can be in a form of video clips, flash games, books, software, images or just text messages. The final goal is to create content that will most probably be shared from the end users in short period of time. Some effective strategies of viral marketing includes: free products, discounts and other strategies that can deliver late profits.

III. Social networks research in R. Macedonia

Some of the related work on social media research in R. Macedonia includes "Online market in Macedonia" [8], "Ipsos Strategic Puls" [9] and "Httpool Macedonia" [10] in July 2010. Their research made in 2010 states that 63% of the responders were using computer with Internet penetration of 53%. According demographic data, 56.4% of responders were male, and according to age groups, users in age group from 15 to 19 are 16%, and smallest group are the age group of over 60 with only 5%. Significant 25% from Internet users are in the age group from 35 to 45 years old, and almost 60% of the responders in RM are in the age group up to 35 years old.

Related research on social media conducted on the territory of R. Macedonia from "Universal Media Skopje" [11] in 2012. Their results are published in "Wave 6 - The Business of Social" [12]. The investigation covers 43% of the world Internet population, with 62 countries included and 41.738 responders. In 2012, first time and Macedonia was part of this world larges research on social media. The investigation of Universal Media is on the effect of the social media on today's global market, and analyses how users in these 62 countries use sites such as Facebook, Twitter for communication with brands, and their expectations. The results are showing that responders from Macedonia in major part are sharing the same habits and opinions with responders from rest of the world. Macedonians on Internet mostly manage some social network profile: 87% have done this in the last 6 months, 84% answered that they have visited some official web page of some company, and 74% joined some on-line community owned by some brand. Macedonians spent almost equal time watching TV and browsing the Internet and social sites. Same as the rest of the world, Macedonian users are concerned about the privacy of the personal information they place on social network sites: 54% answered that they are disturbed with this fact. But, they still expressed their willingness to "sacrifice" part of their personal life to stay in touch with the social events: 30% of responders say that they will miss something if they don't visit their profile on social network sites.

IV. Methodology

The main hypothesis in this paper is that social addons increase the viral effect on some story (news). This hypothesis leads to new hypothesis that usage frequency of social addons is correlated with the time users spend in browsing social network sites. Other questions investigated in this research should confirm/deny some hypothesis including the importance of the age, location, motivation of users engaged in browsing social network sites. Also it investigates the type of content users share using social addons.

A. Hypothesis 1

The time spent browsing social network sites increases, as the age decreases.

B. Hypothesis 2

Social sites are mostly visited from home, with the purpose of keeping a friendship.

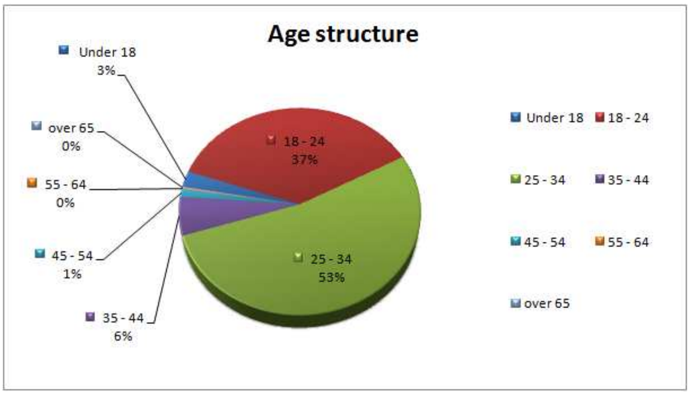


Fig. 2. Age structure of responders.

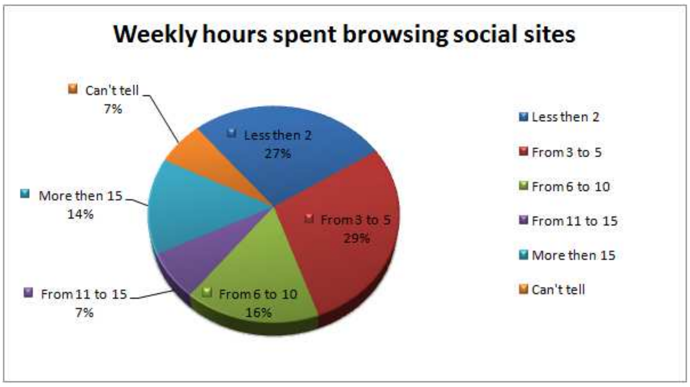


Fig. 3. Weekly hours spent browsing social sites.

C. Hypothesis 3

More time spent browsing social network, leads to more often usage of social addons.

D. Hypothesis 4

Among users of social sites, stories about life and entertainment are more popular than news and sports.

E. Hypothesis 5

By using social addons news are spreading faster (are becoming viral).

F. Users

The research is conducted by surveying two independent groups: internet users (mainly consuming and sharing) and web site or blog owners (mainly offering or aggregating) in Republic of Macedonia. 565 internet users and 24 owners responded on the survey. Largest percent of the responders are in the age group 25-34 years old, with college degree, and 50% from both gender. The owners of the web sites were mainly with beginners experience with social network sites and integration of social addons.

G. Process

The research was conducted in September 2013 and lasted 3 weeks. Two different questionnaires were used In the survey, one for the internet users, and the other for the web site owners. The questionnaires were created using Google Forms.

V. Results

A. Results from Internet users

First questionnaire has total 565 responders, 50.4% male and 49.6% female.

Age structure of the responders is shown on figure 2. Most of the responders (90%) are in the age groups from 25 to 34 and from 18 to 24 years old. Major part of the responders (61%) are with college degree, 26% are with masters degree, and the rest are either high school or Phd.

Regarding the frequency of browsing social sites, by weekly hours spent browsing, the results are shown of figure 3. These results are showing that, 56% from the responders are spending less than 5 hours weekly on social sites, and 37% are spending more then 5 hours weekly. If we compute the mean value of hours spent weekly, for each age group, then the result is 6.2 hours spent weekly browsing social sites.

Responses on the question about the location used browsing social sites, none of the responders answered school as location. Highest percent (85%) of the responders are browsing social sites from home, and this confirms the Hypothesis 2.

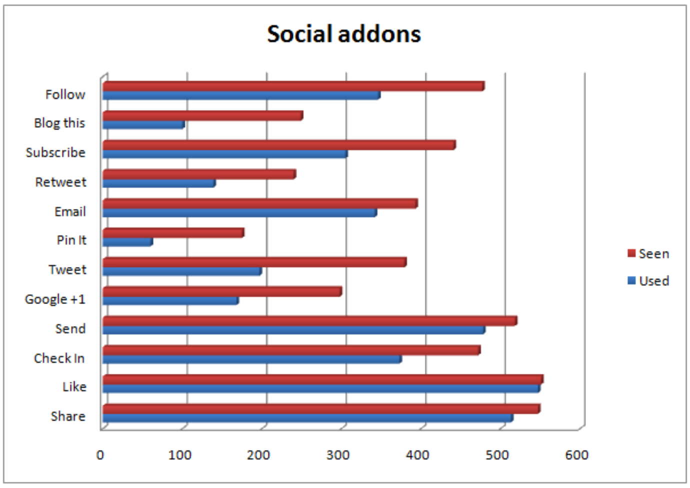


Fig. 4. Social addons.

Next questions are about social addons. The first asks if responders have seen the social addon, and the second one asks if they have used it. Results are shown on figure 4 and we can see that most used social addon is *Like*, where the ratio of seen/used is 99%, and right next to it are *Share* (94%), *Send* (92%) and *Email* (87%). Least used is the social addon *Pin it* (34%), followed by *Blog this* (40%) and *Tweet* (53%).

If we analyze only the results from the responders who visit social networks every day or multiple times a day, then the ration between seen and used is increased significantly. These results are confirming hypothesis 3. More time spent browsing social network, leads to more often usage of social addons.

If we take in to account the gender of responders, men are using more *Subscribe*, *Retweet*, *Tweet* and *Google+1*, and women are using more *Email*, *Pin it*, *Send*. And if we analyze according to age groups, the group over 45 mostly use the social addons *Follow*, *Retweet*, while the youngest group mostly use the social addon *Pin it*. If we take in to account the education level, then the usage of addon *Pin it* increases with the level of education, and same happens for the social addons *Tweet* and *Retweet*.

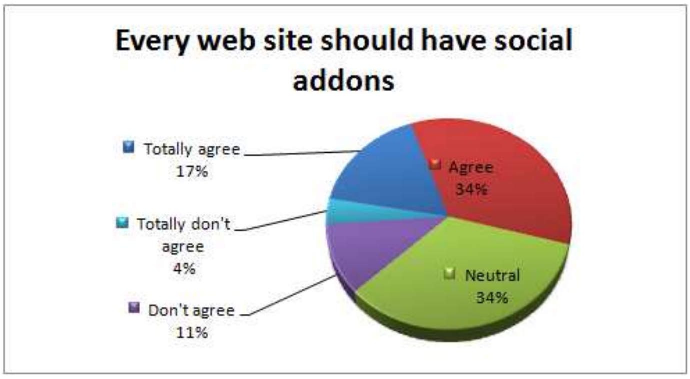


Fig. 5. Every web site should have social addons.

On figure 5 is shown what the responders think of the statement that every web site should have social addons. More than half of them think are agreeing with the statement, but significant part have neutral opinion. Roughly 15% percent of responders do not agree with this statement. Most of these responders who do not agree are in the age group over 45 years old.

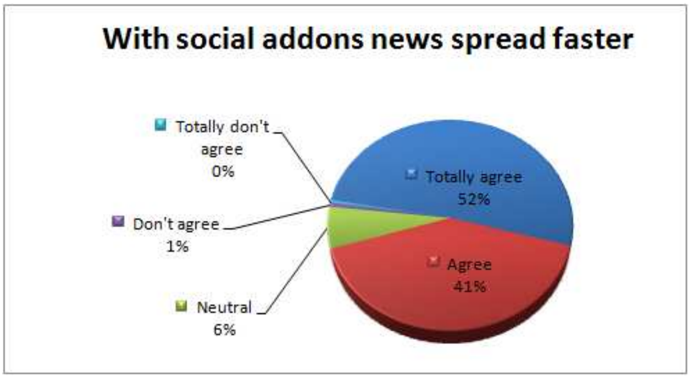


Fig. 6. With social addons news spread faster.

Next statement (figure 6) is about the speed of news spreading (viral effect). Over 90% of the responders are agree with this statement that confirms hypothesis 5. The group over 45 doesn't have a single member responders that does not agree with this statement. Largest percent of the responders who do not agree with this statement are in the age groups of up to 25 and from 25 to 45 years old, while those who totally do not agree with this statement are in the age group up to 25 years old.

B. Results from web site owners

Second questionnaire has total 24 responders. Answers are from owners of web sites or blogs in Republic of Macedonia. Regarding the year their site launched, more then 70% are from the period from 2010 to 2013. There is a single web site launched in 1993, 2000, 2003, 2006 and 2007. On the question what is their target audience, the answers are diverse. There are web sites focused on men, but also there are sites focused only on women. Some of them answered that their main target is the younger generation, only supported by audience of parents, educational workers, city organizations.

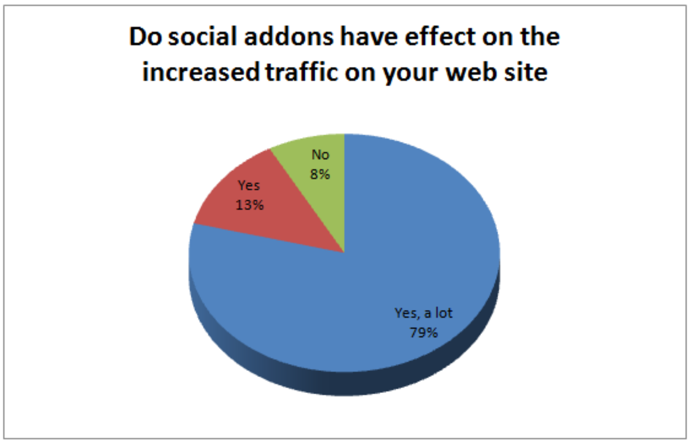


Fig. 7. Do social addons have effect on the increased traffic on your web site.

On the question if the content of their site is automatically published to some social network site, responders answered that they have that implemented only on the following three social network sites: Facebook (46%), Twitter (42%) and Google+ (8%).

On the question if social addons have effect on the increased traffic on their web site, results from answers are shown on figure 7. Almost 80% of the responders answered that social addons have very big effect on the increased traffic to their site, 13% answered that they have small effect, and only 8% answered the opposite. These results are confirming hypothesis 5. Those who answered that social addons do not effect the increased traffic, are using social addons from the beginning and they spend more than 10 hours weekly in promotion of their web site on social networks. They have integrated social addons from Facebook, Twitter and Google+ and the content shared from their web sites is from diverse categories, mostly once per hour or at least once per day.

Those who answered that social addons have effect on the increased traffic to their web site, have integrated all the social addons, and they started circa 2012 and most of them are spending from 6 to 10 hours, or more then 15 hours weekly in promoting their web site on social network sites.

VI. Conclusion

Results from the survey are confirming most of the presented hypothesis. The results for the age structure doesn't conclude the first hypothesis, mostly because the small involvement of the younger age group in the total respondents. The second hypothesis about the location of the users browsing social sites is confirmed, and the results are also showing that friendship is not the dominant motivation for browsing social sites, the fun is stated as most answered choice.

Users spending most hours weekly using social networks, are most often users of social addons, confirming the third hypothesis. The fourth hypothesis concerned with the type of content shared is also confirmed, since life style and fun stories are most shared stories. The fifth and final hypothesis is confirmed both from users and from owners.

References

[1] I.-H. Ting, H.-J. Wu, T.-H. Ho, *et al.*, *Mining and Analyzing Social Networks*. Springer, 2010.

[2] Y. Chen, "Usability analysis on online social networks for the elderly," *Helsinki University of Thechnology*, 2009.

[3] N. B. Ellison *et al.*, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[4] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pp. 49–62, ACM, 2009.

[5] K.-Y. Lin and H.-P. Lu, "Why people use social networking sites: An empirical study integrating network externalities and motivation theory," *Computers in Human Behavior*, vol. 27, no. 3, pp. 1152–1161, 2011.

[6] O. Kwon and Y. Wen, "An empirical study of the factors affecting social network service use," *Computers in Human Behavior*, vol. 26, no. 2, pp. 254–263, 2010.

[7] F. Ellis-Chadwick, R. Mayer, K. Johnston, and D. Chaffey, *Internet marketing: strategy, implementation and practice*. Pearson Education, 2009.

[8] "Online market in macedonia - habits and demand." http://static.httppool.com.mk/Soopshtenie/Newsletter2010.pdf. Accessed: 2013-08-10.

[9] "Ipsos in macedonia (2013)." www.ipsos.com/Country_Profile_Macedonia. Accessed: 2013-09-02.

[10] "Http pool macedonia (2012)." www.httpool.com.mk. Accessed: 2013-09-02.

[11] "Universal media macedonia (2012)." www.universalmedia.com.mk. Accessed: 2013-09-02.

[12] "Wave 6 the business of social." www.universalmedia.com.mk/wave6.html. Accessed: 2013-09-20.

Session 3

Computer Architecture and Parallel Processing, Wireless and Mobile Computing

Next Generation Wireless Technology

The 802.11ac Wireless Network

Davor Naumoski, Biljana Stojcevska, Zlatko Ivanovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
davor.n@bss.com.mk,{stojcevska, zivanovski}@uacs.edu.mk

Abstract - **802.11ac is a fifth generation WLAN standard. The transfer speed of this standard is three times higher than the speed of its predecessor which is much faster than the first standard that emerged in 1997. Its reliability is improved, and both the capacity and the quality of this standard are significantly enhanced. Those features play a key role in its superiority. Its speed reaches up to 1,35Gbps, enabling high bandwidth transfer on multiple devices at the same time. The 5GHz channel used by 802.11ac is less prone to interference. There is a large number of wireless routers that broadcast equivalent omnidirectional signals but a wireless router that is using the 802.11ac standard directs the signal at the location of the 802.11ac wireless device in the network and provides much better network performance. The aim of this paper is to review the characteristics pointed up above and explore them in a device that implements the 802.11ac standard.**

Keywords: 802.11ac, WLAN standards, wireless networks

I. INTRODUCTION

The first wireless standard was 802.11 which was released in 1997 by IEEE [4]. The standard provided speeds of 1Mbps and 2Mbs to be used in office buildings where mobility was not a prerequisite. Since then, the demand for higher speeds and greater coverage is constantly increasing [5].

The appearance of the first commercial product using this technology paved the way to a new era. The benefits of using network connection without physical connection were tremendous leading to rapid spread among the users. With the burst of the number of the devices that use wireless the demand for better performances seems limitless [8]. Over the years Wi-Fi became present at the notebook computers, desktop computers, tablets, TV sets, gaming consoles and smartphones. The main reason why Wi-Fi became such a successful technology is the fact that all new versions are compatible with the previous ones [3].

The latest IEEE standard is 802.11ac which is an improved version of 802.11n. 802.11ac is the first standard that has capability to provide gigabit network access. The speed that can be achieved is 1,3Gbps, which is three times more than the maximum speed provided by 802.11n. The bandwidth is expanded and is 80MHz wide which gives more space for data transfer [1].

The number of new wireless devices that are connected to the Internet is continuously growing and it is expected that the traffic generated by the wireless devices will exceed the traffic generated by the wired devices by the end of year 2014 [9].

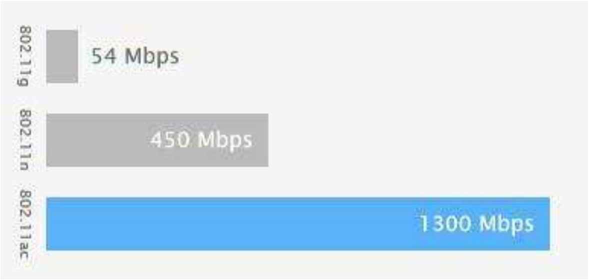


Fig. 1. Wi-Fi data rate

The main aim for 802.11ac technology is to increase the wireless speed and provide gigabit network throughput. For that purpose, 802.11ac introduces a set of improvements at the MAC and the physical layers.

The improvements at the physical layers include:

- increase of the bandwidth per channel;
- increase of the number of spatial streams;
- usage of 256 quadrature amplitude modulation;
- Multiuser Multiple Input Multiple Output at eight spatial streams.

An additional advancement relative to 802.11n is the increased size of the aggregate MAC Protocol Data Units [7].

II. THE USE OF THE WIRELESS DEVICES IN THE HOMES

The evolution of the wireless networks led to a huge increase of the number of the wireless devices in the homes, creating demand for greater Internet access speed [6].

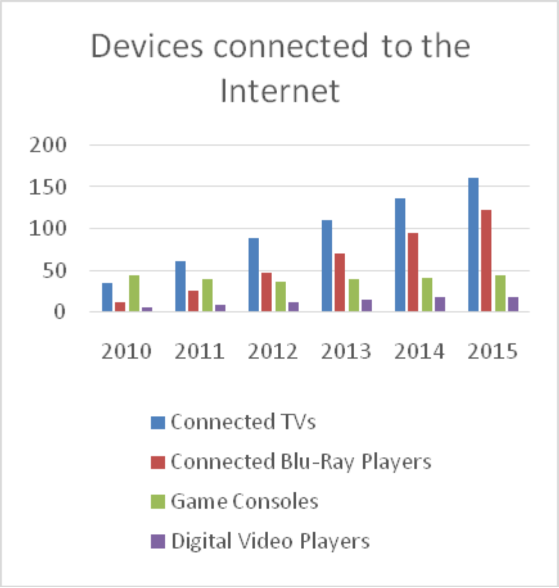


Fig. 2. Millions of devices connected to the Internet

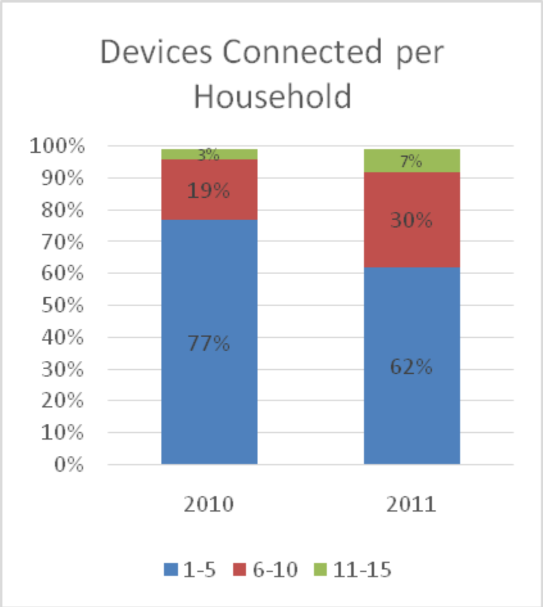


Fig. 3. Number of wireless devices connected per household

The large number of computers, tablets and smartphones in the households that use applications with high-resolution video request networks with high performance and reliability.

The 802.11ac wireless standard is created to satisfy the rising needs for increased performance brought by the increase of the number of Wi-Fi devices.

III. The 802.11ac Wireless Standard

A. 802.11ac speed

The maximum speed that a 802.11n device can reach is 450 Mbps at close range, and the performance decreases with the increase of the distance between the devices. 802.11ac can achieve 1,35Gbps [6]. 802.11ac can maintain higher level of performance than its precursors. This increase is achieved with using wider bandwidth, faster processing and multiple antennas.

B. 802.11ac reliability

Reliability is a significant requirement when high-bandwidth demanding applications are concerned. By increasing the bandwidth volume, 802.11ac offers much greater range for wireless devices which reduces interference and improves reliability. Another improvement is the control of the direction of the broadcasting signal which is named beamforming. While previous standards broadcast the signal in all directions, 802.11ac tracks the relative location of the device and strengthens the signal in the direction of the device [6].

C. 802.11ac quality

802.11b and g devices operate in 2.4 GHz spectrum and use three non-overlapping transmission channels which are used by large number of devices. As a result, the interference is increased while the performances are significantly decreased. The 802.11ac standard operates in the 5 GHz wireless spectrum that is less susceptible to interference. The 5 GHz channel has 23 non overlapping channels which is 8 times more than the capabilities of the 2,4 GHz spectrum. This makes the standard more adequate for applications that are influenced by packet loss and delay [6].

IV. 802.11AC Devices

The fast changes on the market push producers in a battle for new technologies. There are many companies that had recently implemented the new 802.11ac standard in their products. In the next section a representative wireless router that uses this standard will be presented.

A. Apple AirPort Extreme router

AirPort Extreme is a simultaneous dual band 802.11ac wireless router that has the capability to transmit on the 5 GHz band of 802.11a and also the 2.4 GHz band used by the previous standards. Because of this feature, this router can support wireless devices that use 802.11a and also the wireless devices that use 802.11b, 802.11g, and 802.11n at the same time. According to the band that is used by the wireless device the router will automatically establish connection with the band that provides best available performance [1].

Fig. 4. AirPort Extreme

With the development of the 802.11ac standard the producers developed new more powerful antennas that would take all advantages of the new technology. Most Wi-Fi devices transmit equal signal in all directions. AirPort Extreme uses multiple beam forming antennas to track the location of the device and then direct the signal to the device. The effect is a much better signal which provides increased throughput and network range [1].

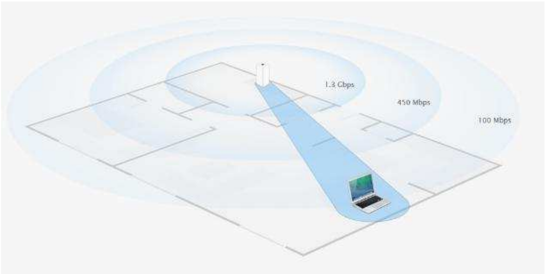


Fig. 5. Beamforming puts the focus on the device

B. Compatibility and security of AirPort Extreme

AirPort Extreme incorporates integrated firewall which protects the wireless network against threats from the public Internet. When AirPort Extreme is installed on the network the firewall is automatically enabled without a need to be specially configured.

AirPort Extreme also provides a Guest Networking function by which a special guest Wi-Fi network can be created where each guest can have its own login credentials. This network allows just Internet access, keeping the internal LAN secure [1].

V. CONCLUSION

The ever-increasing development of the information technologies imposes use of new devices to everyday users which have much better performance characteristics than the older ones. The devices which provide better performance improve everyday life so there is little doubt that in the next few years all the new Wi-Fi devices will implement the 802.11ac standard.

REFERENCES

[1] Apple Inc., AirPortExtreme, 2013, available at: http://www.apple.com/airport-extreme, [last accessed on 27 December 2013].

[2] Apple Inc., Apple AirPort Extreme, AirPort Time Capsule Refreshed, 2013, available at: http://www.technobuffalo.com/2013/06/10/apple-airport-extreme-airport-time-capsule-refresh/, [last accessed on 17 January 2014].

[3] Aruba Networks Inc., 802.11ac In-Depth, 2012, available at: http://www.arubanetworks.com/pdf/technology/whitepapers/WP_80211acInDepth.pdf, [last accessed on 27 December 2013].

[4] IEEE standard 802.11, Part 11: Wireless LAN Medium Access Control and Physical Layer Specifications, 1997.

[5] Motorola Solutions Inc., What you need to know about 802.11AC, 2012, available at: http://www.motorolasolutions.com/web/Business/_Documents/White%20Paper/_Static%20files/80211ac_White_Paper_0712-web.pdf, [last accessed on 8 January 2014].

[6] Netgear Inc., Next Generation Gigabit WiFi - 802.11ac, 2012, available at: www.netgear.com/landing/80211ac/images/wp_netgear_802_11ac_wifi.pdf, [last accessed on 26 December 2013].

[7] Principal Engineer Quantenna Communications Inc., An Introduction to 802.11ac, 2011, available at: http://www.quantenna.com/pdf/Intro80211ac.pdf, [last accessed on 26 December 2013].

[8] Terry A.Francois Cisco Systems Inc., 802.11ac Migration Guide, 2012, available at: https://meraki.cisco.com/lib/pdf/meraki_whitepaper_80211ac_migration.pdf, [last accessed on 10 January 2014].

[9] Xirrus Inc., Building High Performance Networks with 802.11ac, 2012, available at: http://www.xirrus.com/cdn/pdf/building-high-performance-networks-with-802-11ac, [last accessed on 15 January 2014]

Vertical Handover Decision Algorithms:

Current Status and Possibilities

Dejan Svenchev
Marshal Tito 22
Bogdanci, Macedonia
dejan_svencev@yahoo.com

Sonja Filiposka
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
sonja.filiposka@finki.ukim.mk

Abstract—**Achieving seamless handover between different wireless technologies, known as vertical handover, is a major challenge for the next generation wireless networks. In order to achieve seamless continuous connection and select the best network, it is vital to have a good decision making algorithm. In this paper an extensive review and cross comparison of categorized vertical handover decision algorithms are presented.**

Keywords—Heterogeneous networks, Media Independent Handover, Vertical handover, Decision algorithm

I. Introduction

The evolution of mobile technology has increased the demand for multimedia services on mobile devices. Today's popular mobile devices are equipped with several integrated wired and wireless network interfaces. Smartphones are supporting communications through both cellular technologies and Wireless LANs (WLANs); laptops typically come with built-in Ethernet, Wi-Fi, and Bluetooth; tablets include cellular technologies, Wi-Fi and Bluetooth. On the other hand, key success factors for cellular third-generation (3G) communications include better cell capacities, increased data rates, transparent mobility within large geographical areas, and global reachability. As the number of multi-access devices increase, we move closer to a network environment that is often named as "*beyond 3G*" (B3G) representing the combination of cellular networks, metropolitan area networks, wireless local area networks, and personal area networks, i.e. heterogeneous network environment.

Handover management is an important process for mobile networks and becomes imperative in the beyond 3G environment. It enables mobile terminals maintain active connections while moving from one to another network connection. Depending on the network types involved, the handover process can be either horizontal or vertical [1]. A *horizontal* handover occurs between points of attachment that belong to the same network type, while a *vertical* handover takes place between points of attachment supporting different network types. Handovers are also classified in two groups according to the manner of establishing a new connection. A handover is *hard* if the mobile terminal releases its existing connection before connecting to a new base station. A mobile terminal maintaining the current connection till association with a new base station is treated as *soft* handover.

Based on the vertical handover, in order to exploit the vast device capabilities, the IEEE has proposed a new standard: 802.21 Media Independent Handover (MIH) [2], whose main purpose is to enable handovers between heterogeneous technologies without service interruption i.e. to provide session continuity, regardless of the specifics of each technology. A seamless vertical handover between heterogeneous networks without interruption of the on-going services is very important in next generation wireless networks aiming at provisioning of uninterrupted network connections anywhere, anytime.

One of the most important concepts in a heterogeneous networks environment is the Always Best Connected (ABC) [3]. This concept allows connectivity to applications using the devices and access technologies that best suit the user needs, anywhere at anytime. To answer the ABC requirement, various vertical handover decision strategies and algorithms have been proposed recently all designed so as to provide the required Quality of Service (QoS) to a wide range of applications.

In this paper we give an overview of the most interesting and recent algorithms for media independent vertical handover decisions (VHD). The reviews and comparisons made are based on several different classifications, attributes and criteria. Every algorithm weighs its vertical handoff decision based on different attributes. We review a classification of the algorithms in different categories, based on the main handover decision criterion used. For each of the categories, a review of the main characteristics is also presented.

II. Vertical Handover Decision

Every vertical handover decision strategy consists of two stages: Identifying the handover decision criteria and choosing the decision policy to be implemented.

A. *Vertical Handover Decision Criteria*

Handover criteria [4] are the properties that are measured for an indication of the necessity of handover. They are used for choosing the best network and can be grouped as follows:

- Network-related: coverage, bandwidth, latency, link quality (RSS (Received Signal Strength), CIR (Carrier-to-Interfaces Ration), SIR (Signal-to-Interfaces Ration), BER (Bit Error Rate)), cost, security level, etc.
- Terminal-related: velocity, battery power, location information, etc.

- User-related: user profile and preferences.
- Service-related: service capabilities, QoS, etc.

B. Handover Decision Policy

Handover decision policy determines when and where the handover occurs. It represents the network impact on the handover process. Handover decision policy uses the decision criteria to choose the best network by considering the performance of the handover decision. The handover decision policy is mainly concerned about the consequences of the handoff decision such as frequency of handover, latency induced by handover, packet loss during handover, and overall quality of service after the handover.

III. Vertical Handover Decision Strategies

In this section various vertical handover strategies are depicted while emphasizing pros and cons of each one. We can group them as follows: decision function (DF), user-centric (UC), fuzzy logic and neural network (FL/NN), multiple attribute decision (MAD) and context-aware strategies (CA).

A. DF

Decision function is a measurement of the benefit gained by the handover execution. It is calculated for every network *n* that is in range of the user. Every parameter in the decision criteria is assigned a weighted function. So the value for decision function for a network is represented as a sum of all weighted functions of the specific parameters.

The general form of a decision function f_n for service *s* of network *n* with *z* services can be given as [5]:

$$f_n = \sum_{s=1}^{z} \sum_{p=1}^{m} w_{s,p} \cdot c_p^n \qquad (1)$$

where $w_{s,p}$ is the weight assigned to parameter *p* for service *s*, ($\sum_{p=1}^{m} w_{s,p} = 1$) and $c^n{}_p$ is the cost obtained for parameter *p* for service *s* in network *n*. The number of services in an application varies from 1 to *z* and number of parameters for each service varies from 1 to *m*.

H. Wang et al. [6] introduced the first cost function to select the best available network during handover defined as:

$$f_n = w_b \cdot N(1/B_n) + w_p \cdot N(P_n) + w_c \cdot N(C_n) \qquad (2)$$

where B_n is bandwidth of the network *n*, P_n is power consumption of network device for using the network *n*, C_n is monetary cost for using network *n*, $N(i)$ is normalization function of parameter *i*, w_b is weight assigned to bandwidth, w_p is weight assigned to power consumption, w_c is weight assigned to cost. The network with the lowest cost function is chosen as the target network.

The presented decision function can be efficiently used for dynamic network conditions of the network. The policy includes stability period i.e. waiting time to determine whether it is worthwhile to execute the handover. The system separates the decision-making module and the handover module, while offering higher flexibility. To achieve seamlessness, the system considers user involvement with minimal user interaction. Network load balancing can improve the performance of the handover mechanism by preventing many mobile terminals to handover at the same time over the same network. Thus, a performance agent, which collects bandwidth information from the base stations and periodically announce this information to the mobile terminals needs to be implemented.

Chen et al. [7] proposed an adaptive schema based on a utility function, where higher utility leads to target network. The utility function is used to evaluate the discovered available networks, based on factors such as bandwidth and movement speed. Two adaptive handover decision methods are proposed for adjusting the stability period [6] according to the network resources and the running applications on the mobile terminal. The first one calculates utility ratios of the current and the target network, while the latter uses the calculated utility ratios to determine the best network for handover. These two methods reduce the number of handovers and increase the speed of the handover process. An adaptive interface is also proposed for the two adaptive methods. This interface is activated based on the distance between the mobile terminal and the base station with the assistance of location service server. Thus the system discovery method can balance the power consumption and system discovery time.

Major drawback of decision function strategies is that they are not flexible, thus scalability is very low. These strategies fail upright in handling imprecise data. They also do not consider device properties and application demands and are not aware of mobile and network contexts.

B. UC

Vertical handover decision algorithms with user-centric strategies take into account user preferences for user satisfaction. User preferences are represented through cost and QoS as primary factors.

A. Calvagna et al. [8] have proposed two handover decision policies based on threshold value between GPRS and Wi-Fi networks. According to first one the mobile terminal abandons GPRS connection if there is connection interruption. This policy will satisfy the users whose priority is QoS, not cost. In the latter, only Wi-Fi access points with connection blackouts are searched for mobile user's connections. This policy will satisfy the users in terms of connection cost, but they will be frustrated with the QoS. Thus, a balance between QoS and cost is needed, because while the performance of some applications improves, others become worse. To find optimum performance, the cost function can be defined as:

$$C = T_{GPRS} \cdot C_{GPRS} + T_{WiFi} \cdot C_{WiFi} \qquad (3)$$

where T_{GPRS} is time spent by the user over GPRS connection, C_{GPRS} is cost per unit of time over a GPRS connection, T_{WIFI} is time spent by the user over Wi-Fi connection, C_{WIFI} is cost per unit of time over a Wi-Fi connection, and *C* is total cost generated for a given communication session.

The proposed method contains three modules: network selection process module, network monitoring module and user-centric module. Network selection process module contains the decision scheme. Network monitoring module and user-centric module are responsible for reporting network and user preferences related information respectively to the network selection process module. QoS priority policy and cost priority policy both are part of network selection process module. The proposed model was implemented with the help of a distributed mobility protocol, which supports roaming of mobile terminals on GPRS – WIFI integrated platforms.

C. MAD

Multiple attribute decision making (MADM) is handover decision problem that is concerned with choosing between a number of candidate networks with respect to different criteria [9]. MADM deals with the problem of choosing an alternative or candidates based on a set of attributes. The most important steps in MADM methods are: identifying the objectives and goals to be achieved, identification of all decision alternatives (networks) and any related attributes, specification of preferences and ranking the decision alternatives according to specified preferences. Multi attribute utility theory (MAUT) is used by MADM to formalize a common units assessment and specify the decision maker's preferences for each attribute across its respective units scale. MAUT consists of: defining attributes by which decision objectives will be measured, normalizing the measurement or scale of all attributes across all alternatives and weighting the preferences between those attributes. The most popular MADM methods are given below:

- Simple Additive Weighting (SAW) [10]. The score of a candidate network is computed by summing the weights of all attributes.
- Technique for order preference by similarity to ideal solution (TOPSIS) [11]. The network that is closest to the ideal solution and farthest form the worst case solution is chosen as target network.
- Analytic hierarchy process (AHP) [12]. The network selection problem is broke up into a hierarchy of choices and criteria. Then each sub-problem is assigned a weight value.
- Gray relational analysis (GRA) [13] [14]. Ranks the networks and the network with the highest rank will be chosen for the handover process.
- Multiplicative Exponent Weighting (MEW) [15]. Analogous to SAW, where the multiplication instead of addition is the main difference i.e. weighted product of all attributes is used as a score of candidate network.
- Elimination and Choice Translating Reality (ELECTRE) [16]. Performs pairwise comparisons amongst the candidate networks to select the best one.

The MADM ranking is done by assigning a score to each network, which is calculated based on contributions from each included parameter. Also a set of weights are specified in rank calculation. The weights represent the different levels of importance of a parameter for the decision.

Comparison results show that for conversational and streaming traffic classes, almost the same amount of bandwidth is achieved regardless of the used VHD algorithm. For interactive and background traffic classes more bandwidth is obtained when GRA is used. Regarding delay, GRA, SAW, TOPSIS and MEW have similar results for conversational and streaming classes. GRA offers best delay performances in interactive and background classes. When the weight of the jitter increases, all algorithms select the network with the lowest jitter in conversational and streaming classes. And when the weight of BER increases, also all algorithms choose the network with the lowest BER value in interactive and background classes. For all four traffic classes, MEW, SAW and TOPSIS show similar performance, and GRA provides slightly higher bandwidth and lower delay for interactive and background traffic classes. GRA and MEW are very sensitive to the assignment of weights, hence it impacts on their performance.

Due to computational complexity several limitations of MADM methods arise: low speed during the handover, high-ranking time of the available networks. Comparisons indicate that SAW and GRA are the best options in terms of computational complexity.

D. FL/NN

Fuzzy logic is used in situations where decision criteria can contain imprecise information. In one fuzzy inference system, the inputs and outputs to the system should be defined and also membership functions to the input and output variables need to be assigned. The idea is to include expert system in order to control a process whose input-output relationship is described by collection of fuzzy control rules. MADM algorithms can also be conformed for higher performance by fuzzy logic strategies. This is done by combining and evaluating multiple criteria simultaneously in order to develop advanced decision algorithms for non-real time and real time applications.

K. Pahlavan et al. proposed a three-layer back propagation neural network based vertical handover algorithm [17] [18]. The handover process is initiated when there is RSS decrease. The algorithm is intended to satisfy user bandwidth requirements. It was shown that this architecture exhibits better performance than the traditional handover decision algorithms in terms of handover delay and number of redundant handovers. However, too much configuration leads to cost ineffectiveness. Also, the architecture requires preceding knowledge about the radio network.

P. Chan et al. [19] proposed a model that unifies UMTS, GPRS and satellite mobile networks. The goal is to choose the best network for handover based on criteria such as: low cost, good RSS, optimum bandwidth, low network latency, high reliability and long life battery.

E. CA

Context aware handover is based on context information about the mobile terminal and the network [20]. The type of application impacts the decision for the right network, so it is important to find a balance between the requirements of the application and the goals of the user. Context information that is important for handover decision algorithm can be grouped into: terminal-related (its capabilities, location, etc.), user-related (its preferences), network-related (QoS, coverage, etc.) and service-related (QoS requirements, service type such as real-time, interactive or streaming, etc.).

T. Ahmed et al. [21] developed context aware handover decision algorithm that includes session transfer. They take into account mobile initiated and mobile controlled vertical handover decision models. The algorithm is executed for each service type currently running on the device. Primary goals include lowest cost, preferred interface and best quality represented by maximizing throughput, minimizing delay, jitter etc. There are two types of contexts: static and dynamic. The model consists of two pre-configuration stages and three stages of real time calculations:

- *Taking user preferences.* Primary objectives, available interfaces and three types of service are ranked and prioritized.
- *Mapping limit values from discrete preferences.* User preferences are presented as upper and lower bounds for QoS parameters.
- *Assigning scores to available networks.* Reachable networks are evaluated and scored based on preconfigured user preferences.
- *Calculating network ranking.* Each network gets its own rank based on priorities of objectives and scores assigned to the networks.
- *Managing the session.* Includes session transfer scheduling algorithm to switch mobile user application to the new network.

F. Overall Comparison

Multi-criteria choice is essential for vertical handover decision. Traditional and function-based methods consider only a small number of parameters. Adding many constraints may result in reduction of performance on throughput and other network parameters, while the methods become too complicated. So, MADM methods are preferred because more parameters can be used and the problem can be decomposed.

User consideration is also very important in vertical handover decision. It may include user preferences, user interaction or user satisfaction. Under this aspect the most appropriate are user centric and context aware methods.

For *efficiency* of handover algorithm, its goal should be user preference fulfillment. Efficiency means possibility of obtaining a precise decision with good performance. *Flexibility* is defined as separation of the handover process from the whole handover management process and its adaptation with additional parameters. Decision function-based strategies are more flexible for the use of vertical handover policies, but less efficient for real-time applications. For complex problems and imprecise data fuzzy logic and multiple attribute decision algorithms are the best choice. Context aware strategies provide high flexibility as well as high efficiency.

Regarding the *implementation complexity*, neural network based strategy seems to be more complex due to complicated topology and difficult practical implementation. Fuzzy logic and neural network based strategies examine only a few context parameters and can be complex to adapt for multimode mobile terminal with limited resources. On contrary, there are context aware strategies that apply MAD methods that use simpler calculation compared to FL/NN strategies.

According to *service types*, non-real time applications are supported by all the strategies, while real time applications are not supported by traditional and user centric strategies.

Finally, we can conclude that context aware strategies have best overall performance according to the previously given characteristics. Next, good performances are demonstrated by multiple attribute decision and fuzzy logic strategies. Fuzzy logic and context aware strategies are improved in the way that can be combined with multiple attribute decision strategy.

IV. VHD Grouping Based on Handover Criteria

In this section various vertical handover algorithms are grouped based on the main handover criterion used. The main function of each algorithm and its advantages/disadvantages are also depicted. Four categories are considered: RSS based, bandwidth based, cost function based and combinations.

A. RSS Based VHD Algorithms

A number of studies have been realized, due to simplicity of hardware requirements for RSS measurement. To make the handover decision the algorithm compares the RSS of the current point of attachment with the RSS of the candidate's point of attachment.

The schemes for RSS comparison are [22]: relative RSS, RSS with hysteresis and RSS with hysteresis plus dwelling timer. Relative RSS is not used for VHD, since comparison cannot be executed directly on different types of networks due to distinction of involved technologies. In the RSS with hysteresis method, handover is executed when the RSS of a new base station is higher than the RSS of the current base station by a predefined value. In RSS with hysteresis plus dwelling timer method, the timer is set when the RSS of a new base station is higher that the RSS of the current base station. When the timer reaches a certain value, handover is performed. Using this method the ping pong handovers are minimized. Also other network parameters such as bandwidth can be combined with RSS in the VHD process.

Zahran et al. [23] proposed algorithm for handovers between 3G and WLAN networks in which RSS weight is combined with lifetime metric or available bandwidth of WLAN candidate. The lifetime metric represents the expected time period in which the user is able to maintain connection with WLAN. Similarly, Mohanty et al. [24] proposed algorithm for WLAN to 3G handover in which a dynamic threshold is compared with the current RSS. The dynamic threshold provides reduced number of redundant handovers and handover failure will be under a given limit. Yan et al. [25] proposed algorithm that take into regard the time the mobile terminal will stay in a WLAN cell. Thus the handover will happen when this time is greater than the time threshold.

B. Bandwidth Based VHD Algorithms

Bandwidth based algorithms are based on the available bandwidth as main decision criteria. Lee et al. [26] proposed algorithm in which the remaining bandwidth, the state of the mobile terminal and user service requirements are used for handover decision form WLAN to WWAN network and vice versa. Handover to the preferred network is executed if the mobile terminal is in the idle state, else the user application type is a criterion for handover decision. Thus, delay-sensitive applications get lower handover latency. Another similar bandwidth based VHD algorithm that uses SINR is proposed by Yang et al. [27] for handover between WLAN and WCDMA network. It guarantees higher throughput than RSS based handover, because the available throughput relies on the SINR. But the algorithm can cause excessive handovers by SINR variation that leads to the ping-pong effect. Chi et al. [28] devised VHD algorithm that uses WDP (Wrong Decision Probability) prediction. That is combination of the probability of unnecessary and missing handovers. This algorithm reduces the redundant handovers and WDP and balance the traffic load, but does not take into account RSS.

C. *Cost Function Based VHD Algorithms*

These algorithms perform handover decision based on value of the cost function for each network. The cost function is a combination of metrics such as monetary cost, security, power consumption and bandwidth, and have different weights regarding the network conditions and user preferences.

Zhu and McNair [29] developed cost function based algorithm that calculates the cost of the candidate networks i.e. the sum of cost of each QoS parameters (e.g. bandwidth, battery power and delay). First, the algorithm prioritizes all active applications and then the cost for each candidate network is calculated for the service with the highest priority. The handover will occur between the application with highest priority and the network with the lowest cost. The algorithm increases the number of user efficient requests and reduces the handover blocking probability.

Hasswa et al. [30] proposed cost function based algorithm that includes normalization and weight distribution. It calculates the network quality factor to estimate the performance of a candidate network. The handover will be performed if the quality factor of the current network is smaller than the one of the candidate network. Increased system throughput and user satisfaction are the main benefits.

Tawil et al. [31] proposed weighted function based algorithm whose calculation is done in the visited network, instead of the mobile terminal. As a result, the resources of the mobile terminal are conserved and therefore the handover delay is decreased, the handover blocking rate is lowered and the throughput is increased. However, due to communication between the mobile terminal and the point of attachment of the visited network there may appear additional latency and excessive load in a number of mobile terminals.

D. *Combination Algorithms*

Combination algorithms incorporate a number of parameters such as the ones used in cost function algorithms for handover decision. These algorithms are built on artificial neural networks or fuzzy logic.

Nasser et al. [32] proposed VHD algorithms based on artificial neural networks (ANN). The mobile device gathers parameters from the candidate wireless networks like network usage cost, network security, transmission range and network capacity to facilitate the handover decision process. Then these parameters are forwarded to a unit called handover manager. The handover manager includes three components: network handling manager, feature collector and ANN training/selector. The best handover wireless network is selected by a multilayer ANN based on the user's preferences. The algorithm advantage is successful network election by choosing appropriate learning rate and acceptable error value. But the main drawback is the long delay during the training process.

Xia et al. [33] deployed fuzzy logic method that uses fuzzy parameters to deal with handovers between WLAN and UMTS. The input parameters include current RSS, predicted RSS and bandwidth that are used to determine network performance. A pre-decision unit is used in the algorithm. If the mobile terminal is connected to the UMTS and the WLAN is available, the pre-decision unit eliminates redundant handovers when the velocity of the mobile terminal is larger than the velocity threshold. After the pre-decision, fuzzy logic based normalized quantitative decision is used to evaluate the performance of the target networks. The algorithm improves the performance by decreasing the redundant handovers and avoiding ping-pong effect.

E. *Comparison*

Regarding *applicable networking technologies*, RSS based algorithms handle with handovers between 3G and WLAN networks. The other types of algorithms are used for handovers between any two heterogeneous wireless networks. For example bandwidth based algorithms can be applied between WWAN and WLAN networks or between WCDMA and WLAN. The WLAN networks are used due to the high bandwidth and cost efficiency.

According to *input parameters*, in RSS based algorithms the fundamental parameter is RSS. In bandwidth based algorithms usually RSS in combination with bandwidth is used. Cost function based and combination algorithms utilize different network parameters such as monetary cost, bandwidth, security and power consumption. In RSS based and bandwidth based algorithms, RSS and the bandwidth are used as main target selection criteria, respectively. Whereas in combination and cost function based algorithms the target network is the one with the highest overall performance that is derived based on the various network parameters.

In terms of *complexity* the RSS based algorithms are the simplest. These are followed by the bandwidth based algorithms. Cost function based algorithms are more complex due to the collection and the normalization of the various network parameters. As the most complex are considered the combination algorithms, because of the prior system training.

According to *reliability*, RSS variation reduces the reliability of the RSS based algorithms. Also the reliability of bandwidth based algorithms is decreased by obstacle of available bandwidth measurements. In cost function based algorithms the constraints for measuring some parameters such as security level degrade reliability of the algorithms. And the combination algorithms seem to be the most reliable because of their pre-training requirements.

RSS based algorithms stand out with one benefit in common: decreasing the number of handovers. For throughput, bandwidth based and cost function based algorithms are better than others. Combination algorithms have disadvantage in the longest delay among the others because of system complexity.

V. CONCLUSION

Beyond 3G networks aim to provide seamless mobility between different access technologies so that the user can be connected on the best available network. This network usually provides better service at lower cost to the user and improves utilization of the system resources. A vertical handover process combined with well-designed handover algorithms can provide seamless services.

In this paper we introduced a thorough comparison of vertical handover decision algorithms. The main goal of all these algorithms is to find the best available network for handover and the appropriate time to perform handover. Thus choosing the right VHD algorithms that meet the requirements of both user and network providers is very crucial. We presented the handover decision process with a classification of the different VHD algorithms along with their pros and cons. Also handover decision criteria that are needed to perform better handover decision for user satisfaction as well as for the efficient use of the network resource are depicted. The presented comparison allows for choosing the appropriate decision algorithm based on the expected outcomes.

ACKNOWLEDGMENT

This work has been partially supported by the EUROWEB project (http://www.mrtc.mdh.se/euroweb) funded by the Erasmus Mundus Action II programme of the EC.

REFERENCES

[1] N. Nasser, A. Hasswa and H. Hassanein, "Handoffs in fourth generation," IEEE Communications Magazine 44, 2006.

[2] IEEE 802.21 WG, "IEEE Standard for Local and Metropolitan Area Networks. Part 21: Media Independent Handover Services," IEEE Std 802.21-2008, 2009.

[3] E. Gustafsson and A. Jonsson, "Always best connected," IEEE Wireless Communications, 2003.

[4] Mrs.Chandralekha and P. Dr.Kumar Behera, "Minimization of Number of Handoff Using Genetic Algorithm in Heterogeneous Wireless Network," Latest Trends in Computing, 2010.

[5] J. McNair and F. Zhu, "Vertical Handoffs in Fourth-generation Multinetwork Environments," IEEE Wireless Communications, 2004.

[6] H. Wang, R. Katz and J. Giese, "Policy-enabled handoffs across heterogeneous," Second IEEE Workshop on Mobile Computing Systems and Applications, 1999.

[7] W. Chen, J. Liu and H. Huang, "An adaptive scheme for vertical handoff in wireless overlay networks," Proceedings on the 10th International Conference on Parallel and Distributed Systems, 2004.

[8] A. Calvagna and G. D. Modica, "A user-centric analysis of vertical handovers," Proceedings of the Second ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots, 2004.

[9] R. Ribeiro, "Fuzzy multiple attribute decision making: a review and new preference elicitation techniques," Fuzzy Sets and Systems, 1996.

[10] W. Zhang, "Handover Decision Using Fuzzy MADM in Heterogeneous," Proc. IEEE WCNC'04, 2004.

[11] L. Sheng-Mei, P. Su and X. Ming-Hai, "An Improved TOPSIS Vertical Handoff Algorithm for Heterogeneous Wireless Networks," IEEE International Conference on Communication Technology (ICCT), 2009.

[12] T. Saaty, "How to make a decision: the analytic hierarchy process," European Journal of Operational Research, 1990.

[13] K. Yoon and C. Hwang, "Multiple Attribute Decision Making: An Introduction," Sage Publications, 1995.

[14] K. Savitha and C. Chandrasekar, "Grey Relation Analysis for Vertical Handover Decision Schemes in Heterogeneous Wireless Networks," European Journal of Scientific Research, 2011.

[15] Q. Song and A. Jamalipour, "A Network Selection Mechanism for Next Generation Networks," Proc. IEEE ICC'05, 2005.

[16] F. Bari and V. Leung, "Application of ELECTRE to network selection in a heterogeneous wireless network environment," IEEE WCNC, 2007.

[17] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. Makela, R. Pichna and J. Vallstron, "Handoff in hybrid mobile data networks," IEEE Personal Communications 7, 2000.

[18] J. Makela, M. Ylianttila and K. Pahlavan, "Handoff decision in multiservice networks," The 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2000.

[19] P. Chan, Y. Hu and R. Sheriff, "Implementation of fuzzy multiple objective decision making algorithm in a heterogeneous mobile environment," IEEE WCNC, 2002.

[20] Q. Wei, K. Farkas, C. Prehofer, P. Mendes and B. Plattner, "Context-aware handover using active network technology," 2006.

[21] T. Ahmed, K. Kyamakya and M. Ludwig, "A context-aware vertical handover decision algorithm for multimode mobile terminals and its performance," Proceedings of the IEEE/ACM Euro American Conference on Telematics and Information Systems, 2006.

[22] G. Pollini, "Trends in handover design," IEEE Communications Magazine, 1996.

[23] A. Zahran, B. Liang and A. Saleh, "Signal threshold adaptation for vertical handoff in heterogeneous wireless networks," Mobile Networks and Applications, 2006.

[24] S. Mohanty and I. Akyildiz, "A cross-layer handoff management protocol for next-generation wireless systems," IEEE Transactions on Mobile Computing, 2006.

[25] X. Yan, N. Mani and Y. Sekercioglu, "A traveling distance prediction based method to minimize unnecessary handovers from cellular networks to WLANs," IEEE Communications Letters 12, 2008.

[26] C. Lee, L. M. Chen, M. Chen and Y. Sun, "A framework of handoffs in wireless overlay networks based on mobile IPv6," IEEE Journal on Selected Areas in Communications 23, 2005.

[27] K. Yang, I. Gondal, B. Qiu and L. Dooley, "Combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks," IEEE GLOBECOM, 2007.

[28] C. Chi, X. Cai, R. Hao and F. Liu, "Modeling and analysis of handover algorithms," IEEE GLOBECOM, 2007.

[29] F. Zhu and J. McNair, "Optimizations for vertical handoff decision algorithms," IEEE WCNC, 2004.

[30] A. Hasswa, N. Nasser, H. Hassanein and Tramcar, "A context-aware cross-layer architecture for next generation heterogeneous wireless networks," Proceedings of the 2006 IEEE International, 2006.

[31] R. Tawil, G. Pujolle and O. Salazar, "A vertical handoff decision scheme in heterogeneous wireless systems," 67th VTC, 2008.

[32] N. Nasser, S. Guizani and E. Al-Masri, "Middleware vertical handoff manager: a neural network-based solution," IEEE ICC, 2007.

[33] L. Xia, L.-G. Jiang and C. He, "A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method," IEEE ICC, 2007.

Differential Evolution and Parallelization

Matjaž Depolli
Department of Communication Systems
Jožef Stefan Institute
Ljubljana, Slovenija

email: matjaz.depolli@ijs.si; fax: +386 1 477 3111

Abstract—**Differential Evolution (DE) is an optimization algorithm, that uses principles form biological evolution to stochastically optimize real-valued functions. As a stochastic algorithm, it requires a large number of iterations (it converges slowly) to reach good quality solutions. Each iteration comprises the cost-function evaluation, over which the DE has no control, it is essentially a black-box accepting input and returning output. Since the cost function evaluation is also usually the most time demanding part of the algorithm, the parallelization principle seems straight forward - have the evaluations execute in parallel. The problems to be solved, though, are the dependence between sequential evaluations and the low efficiency when the cost function can take variable amount of time to complete. In this paper, we tackle the problems by implementing asynchronous master-slave parallelization scheme. We demonstrate how this parallelization scheme solves both presented problems but also causes an unwanted change in the differential evolution behavior. We asses the magnitude of the change and show the parallel algorithm performance on two optimization by simulation examples.**

***Keywords:** parallel, evolutionary algorithms, simulation, optimization*

I. Introduction

Differential Evolution (DE) [1] is an optimization algorithm that uses principles form biological evolution to stochastically optimize real-valued functions. As a member of stochastic optimization algorithms family, it does not guarantee to find the global optimal solution nor to always find the same solution. On the other hand, the requirements of DE regarding the function to be optimized are also very low. Function does not need to be differentiable or even analytically solvable; it can be a complete black-box with no known functionality other than the fact that it takes an input and returns an output, and that the two are somewhat related.

The function to be optimized is called *cost function* or *objective function*. DE is best used on real-valued cost functions although it can solve discrete valued cost functions too. Fro the latter case though, better algorithms can be easily found that are dedicated to discrete functions [2]. DE, as all evolutionary algorithms, works by executing the cost function multiple times using different parameters every time. Executing the cost function on a parameter set and assigning it its cost is called *an evaluation*, since it evaluates the parameters. Although multiple executions of cost function is very similar to the Monte Carlo procedure, DE has an important advantage. The parameters to cost function are not selected entirely on random, but using a heuristic that is inspired by the biological evolution. The heuristic guarantees to guide the algorithm towards the exploration of parts of the solution space which are more likely to contain optimal solutions. Since large number of evaluations are required and the cost function evaluation is also usually the most time demanding part of the algorithm, the optimization using DE can be very time consuming. One possible way of making it faster is by parallelization [3].

Parallelization is used for two main purposes. First is to speedup the data processing [4], and second is to gain the ability to manage more data [5]. Third purpose of parallelization is moving into the spotlight and is the mixture of the main ones – to gain the ability to process massive amounts of data in a reasonable time frame, such as real-time processing of data [6]. This work uses the parallelization strictly to make execution faster.

Parallelization principle for DE is in essence very straight-forward – since the evaluations do not depend on each other, they should execute in parallel. While this holds for the evaluations themselves – several executions of the cost function are independent, it does not hold for the selection of parameters. As previously noted, the parameters for various executions are selected by a heuristic, which depends on the history of already evaluated parameter sets. When some of the evaluations are performed in parallel instead of sequentially, the heuristics has to work with a more limited history. This limitation was discussed in [7], on the grounds of a very similar parallel evolutionary algorithm (AMS-DEMO) and also explained how it applies to other, related algorithms. The principle of parallelization that minimizes this limitation was shown and the performance of the parallel algorithm evaluated. Here, the results of the study are used, including the parallelization principle, and applied it on the simpler DE algorithm.

In this paper, an asynchronous parallelization of DE, using the master-slave principle, is presented. In the rest of the document, this implementation shall be referred to as the *Asynchronous Master-Slave DE* implementation (AMS-DE).

The rest of the document is organised as follows. Section II provides the details of the AMS-DE implementation and its efficiency. We demonstrate the speedups of AMS-DE on two simulation-based optimization problems, which are presented in section III. In section III, the experimental methodology is provided, and in section IV, the results of the experiments are given. The conclusion follows in section V.

II. Asynchronous Master-Slave Differential Evolution

Differential evolution operates with *individuals* or *solutions* – vector encodings of parameters to the cost function. At any given moment the algorithm holds a working set of solutions, called the *population*, in memory. In each iteration, it applies the *evolutionary operators* (crossover, mutation, and recombination) to form a single new candidate solution. The candidate solution is then evaluated and inserted into the population if it performs better than its predecessor. For the next iteration, the newly evaluated solution can be considered by the evolutionary operators for creation of a new candidate. This immediate use of new solutions is called the *steady-state* principle, and DE is an example of steady-state evolutionary algorithms [1]. Steady-state principle promotes fast convergence of solutions towards an optimum on the account of having a very sequential nature.

The master-slave principle splits the task of an evolutionary optimization into two sub-tasks: the application of evolutionary operators and the evaluation of solutions. The former is performed by the master process, which also holds the population in memory, while the latter is performed by slave processes in parallel. When there is no synchronization performed between the master and slaves, other than starting the processes at the start and ending the processes when the master decides the best solution is good enough, then the principle is said to be asynchronous. Asynchronous master-slave DE thus works similarly to DE but with the iteration procedure changed. The principle is explained in detail in [7], so only a short description follows.

In AMS-DE, evaluations are executed on remote slaves, and the call to evaluation on the master is replaced by two independent non-blocking parts – a call to the remote evaluation and a check for pending results from remote evaluations. These parts are independent, therefore the master can issue multiple calls for remote evaluation before receiving a single result. It can also receive several results before issuing another remote evaluation. Since parts are non-blocking, master never waits for any particular result from a slave, but rather creates new candidate solutions and issues them into evaluation as long as there are idle slaves available. When there is no idle slaves, master waits to receive results – evaluated solutions, which it then immediately incorporates into his population. Since the workload of the master process is very different from workloads of slaves, much lower in case where cost function is at least in part simulation, the master process shares the processor with one of the slave processes, maximizing the algorithm parallel efficiency.

III. Problem Cases

A. Optimization of simulated ECG shape

First problem case covers the optimization of simulated ECG shape via parameters of an ECG simulator. The ECG simulator [8] comprises a discretized 3D model of a heart (see Figure 1), an *Action Potential* (AP) shape function [9], and a rule for varying the AP shape across the heart model. The 3D heart model is based on a geometrically simplified shape, resembling the ventricles of the human heart. The discretization creates small cells of 1 cubic millimeter volume. Each cell is assigned its own AP shape, defined by the AP shape function, which is basically a stencil for the cell electrical activity. Simulated ECG is represented as the measurement of an electrical field from a position outside of the heart, where the electrical activities from all of the cells contribute to the total measurement. This outside position is simplified as if it were lying in vacuum, so that the tissue surrounding the heard does not need to be simulated. The optimization procedure is allowed to modify several parameters of the AP shape function, while the rule for varying AP shape across the model is fixed. The ECG simulator takes almost constant time to calculate the ECG and therefore seems very suitable for the parallelization principle of running individual simulator runs (cost-function evaluations) in parallel.

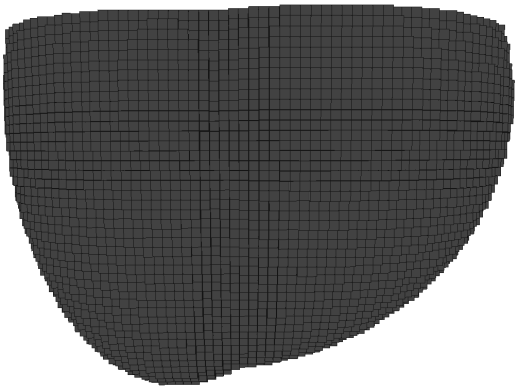


Figure 1: The heart model (only left and right ventricles) composed of 3D cells.

Cost function for the problem takes 8 parameters that determine the AP shape functions. Its return is formed by running the simulator and obtaining the Pearson correlation coefficient between a predefined measured ECG signal and the simulated ECG signal, and subtracting it from 1. This way the cost function is minimal (0) when the ECGs are the same, and maximal (2) when they are inversely correlated. Note that only a part of a single heart beat from ECG signals is simulated and compared – so called T and U waves [8] – since these are known to be controlled by shapes of APs in individual cells. These waves occur roughly between 200 and 700 ms after each heart beat start.

B. Minimization of energy losses in air cavity

Second example tackles with a simulation of a physical device. In a differentially-heated air-filled cavity (see Figure 2), some non-permeable obstacles can be placed. The cavity is cooled on the top and heated on the bottom horizontal edge (Dirichlet boundary conditions), with left and right vertical walls perfectly isolated (Neumann boundary conditions). Such setup creates natural convection – motion of the air, that mixes the cold and warm air and consequently promotes energy transfer from the bottom to top. The energy transfer is considered energy loss of the system. The optimization problem is placing 5 square obstacles in such a way that the energy losses of the cavity will be minimal, or in other terms, the thermal insulation property of cavity will be maximal. Since the air and the obstacles are thermally conductive, the means of minimizing energy losses is only by limiting natural convection.

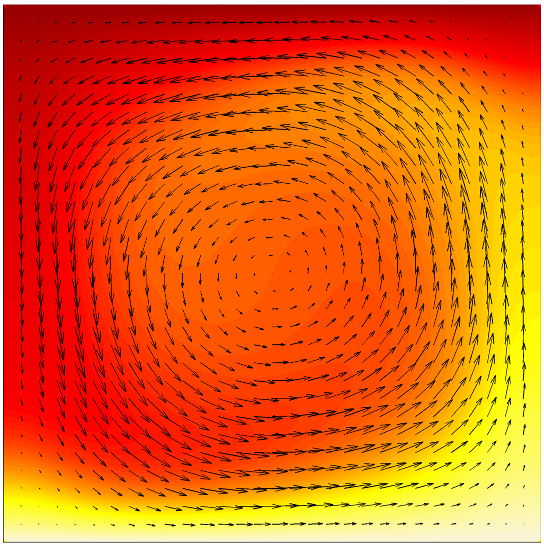


Figure 2: Temperature distribution (colored red – cold to bright yellow – hot) and fluid flow (black overlaying arrows) of an empty cavity. The energy losses are clearly mostly due to thermal convection forming a large vortex that is transporting hot air upwards and cold air downwards.

The simulation is solved with a strong-form local meshless numerical method [10], by a methodology that has been found a reliable and effective tool for solving complex multiphase problems. The basic idea of the method is to use local approximations of the considered physical field to evaluate differential operators. The size and shape of local support, i.e. the subset of computational nodes used in the local approximation, is not restricted in any way. Although there is no restriction regarding the dimension of the system, we use the collocation approach, i.e. the number of support points is the same as the number of basis functions [11]. The temporal discretization is done through a two-level explicit time stepping.

The simulator implementing the numerical method takes variable amount of time to complete the simulation, since it starts form constant initial conditions and stops once the fluid motion and temperature stabilize (less than 0.01% change in absolute temperature between consecutive time steps, for all discretization nodes of the simulation). Some rare cases where the fluid does not stabilize in 20 seconds of simulation time are considered chaotic and marked as infeasible within the optimization procedure. In all feasible cases, the energy flux measured at vertical walls is taken as cost function for the optimization procedure.

Cost function for this case takes 10 parameters – (x, y) coordinate pairs for five obstacles. The domain size is fixed (1 x 1) and obstacles are also of predefined size (0.1 x 0.1). Their positions are allowed to vary between (-0.3, -0.3) and (1.3, 1.3). These ranges clearly allow for the obstacles to fall outside of the domain, which is by design – it gives optimization the option of not using all the obstacles, if they are not needed. In cases of obstacles overlapping each other, simulator considers them as one bigger obstacle with the shape being the union of the overlapping obstacle shapes. The cost function returned is calculated as the heat flux between the Dirichlet boundaries. Note that, due to the energy conservation fluxes on both Dirichlet boundaries are the same.

IV. Methodology

The following parameters are used for the DE algorithm: population size 30, scaling factor 0.2, crossover probability 0.2, DE scheme “rand/1/bin”. The definition of all these parameters can be found in [1].

The experiments were run on a computer cluster, composed of 36 computers, interconnected over a Gigabit Ethernet connection switch. The number of used computers varied between 1 and 24 for the experiments. The cluster could not be reserved for the use of more computers since it was partly occupied by other jobs for the whole time the tests were performed. Since the optimization algorithm is stochastic, each experiment was repeated 5 times to ensure the consistency of results. Each problem case represents its own set of experiments. Absolute running times of individual experiments were measured and the best result (lowest cost) was saved. Since the results obtained are highly variable due to the stochastic nature of the optimization algorithm, all optimizations are not stopped after some predefined solution cost is achieved but rather after a fixed number of performed cost-function evaluations.

Speedup *S* is defined with the following equation:

$$S=\frac{t_n}{t_1} ,$$

where t_n is the execution time of AMS-DE on *n* processors, t_1 is the execution time on *1* processor, and both executions return equivalent solution regarding its cost. Since the stochastic nature of the AMS-DE algorithm, there is a lot of randomization in returned solution quality and there would be a forbidding amount of noise present in the results if the algorithm was required to reach an equivalent solution in each run. Instead, the requirement is lowered to all the runs making the same number of evaluations. This would actually be an equivalent requirement if the convergence rates of all runs were the same.

Unfortunately, a very large number of experiments would be required to confirm that the convergence rate of AMS-DE is not statistically significantly different on different experimented numbers of processors. The rule-of-the-thumb, taken from [7], for the convergence of a parallel algorithm using AMS principle to remain constant is that the population size is larger than the number of slave processes. In the both presented cases the criterion is satisfied and we assume the convergence of AMS-DE does not decrease significantly and therefore the requirements for speedup calculations are satisfied.

V. Results

A. Optimization of simulated ECG shape

Although this paper focuses on the parallelization aspect of the optimization, the optimized simulated ECG shape is presented here. Figure 3 shows the simulated and measured ECG. Their agreement is very good up until about 400 ms into the heart beat, with great timing and general shape of the first peak, but afterwards, the waveform is no longer simulated correctly. Conclusion can be made that the simulator lacks some mechanism for generating ECG properly across the observed time interval but this is out of scope and left for future work.

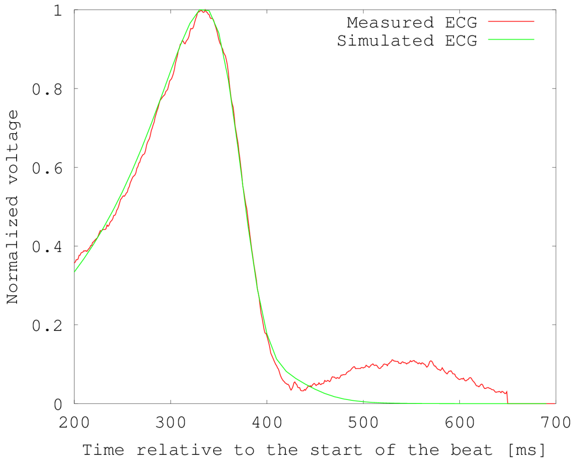


Figure 3: Comparison of the simulated and measured ECG (only ECG waves that are most influenced by AP shape)

Next, the convergence of results is presented. Convergence is deduced from the best solution cost as a function of the number of performed evaluations. Figure 4 shows the convergence of 5 runs. Y axis is shown logarithmically because the changes in solution quality is getting progressively smaller with the optimization procedure converging. This is mostly due to the behavior of the correlation measure that is used for the cost function. Nevertheless, those small improvements at the end are not unimportant. Note that regardless of the number of computers used, the convergence plots appear very similar.

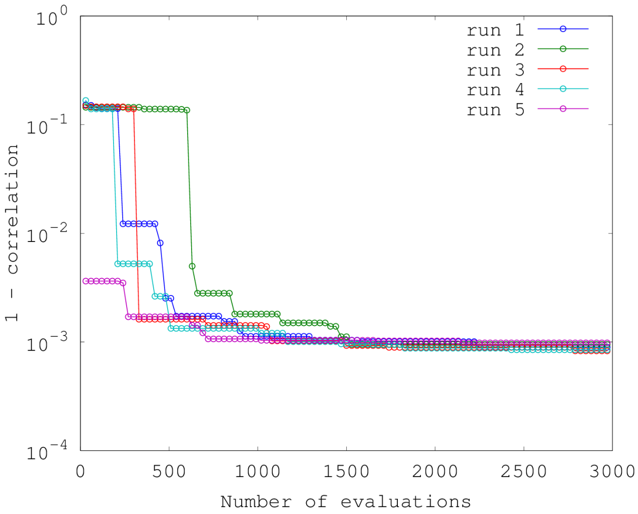


Figure 4: Convergence of solutions for ECG shape optimization of 5 runs on a single computer.

Solutions can never reach cost of 0, because the ECG measurement that serves as the reference is noisy (high frequency noise that makes the measurement appear a bit toothed), while the simulated ECG is clean of noise.

Finally, the speedup is calculated and averaged over 5 runs. Figure 5 shows it is almost linear, as was expected. As mentioned previously, and also seen from the variation of the convergence in Figure 4, a more accurate speedup analysis would require a much larger number of experiments, which was not achievable for this work.

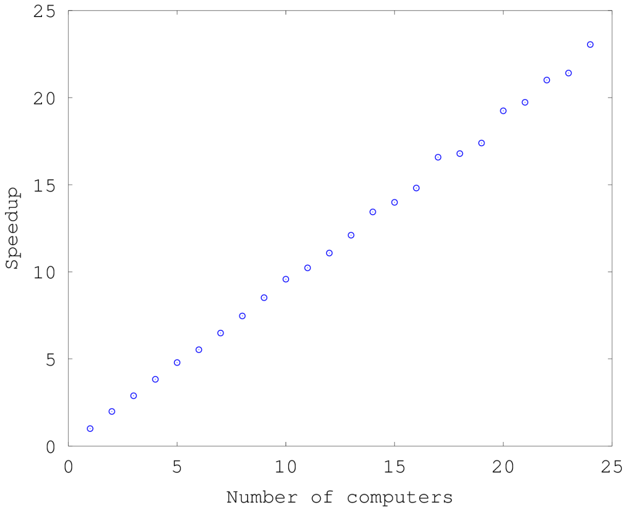


Figure 5: Speedup of the ECG shape optimization on 1 to 24 computers

B. *Minimization of energy losses in air cavity*

First, the general results are presented. Optimal obstacle placement that was found is shown in Figure 6. Similar placements were found in multiple optimization runs that were performed, but most of them forming a related pattern of breaking the natural convection vortex into two smaller vortices close to the middle of the domain. The cost function optimums varied between 2.3 and 2.6 times lowered energy losses (43 to 38 % of energy losses of empty cavity). Clearly the optimization was able to find a principle for minimization but could not fully minimize losses.

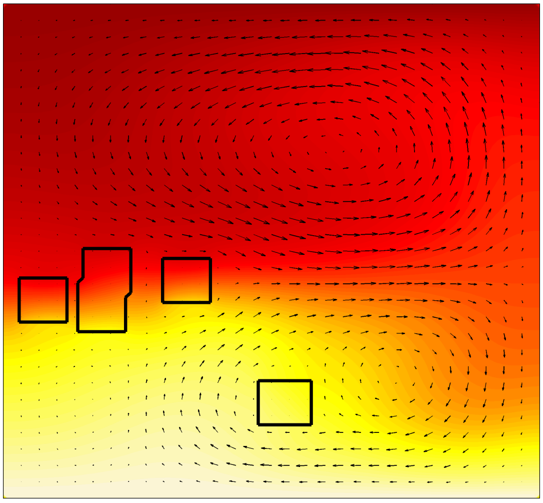


Figure 6: One of the obstacle placements that results in minimal energy losses (2.6 times lower losses than in an empty cavity). Natural convection vortex is limited by splitting the vortex into two and additionally obstructing the bottom vortex. It seems though that the obstacles could be placed in a more optimal manner, but optimization was not able to improve this placement.

Next, the convergence of solutions is checked. Figure 7 shows the convergence of the heat flux towards its optimum for 5 runs. As in previous case, the image does not change significantly no matter which 5 runs are taken into account. Heat flux is displayed relative to the heat flux of empty cavity.

Unlike in the previous case, optimization starts with relatively good solutions from initial (random) population and then improves them by a small amount. Random solutions work well in this case since even randomly placed obstacles manage to greatly disrupt the natural convection. Optimization is still useful since it finds better patterns of placement. It does appears though, that given these patterns as a starting point, local optimization would be able to optimize them further.

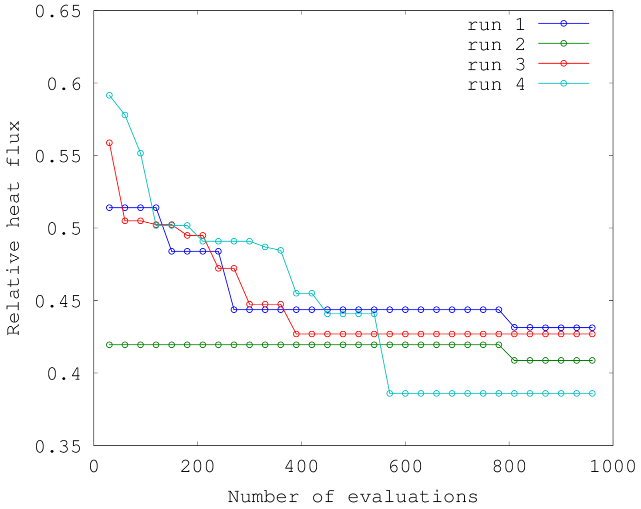


Figure 7: Convergence of solutions for energy losses minimization of 5 runs on a single computer.

Finally, speedup as the function of the number of processors is shown in Figure 8. As in the first case, speedup is nearly linear. In this case, however, it is much more jagged, due to higher variability of the simulator-based cost function. Again, a much higher number of repetitions would have given a smoother picture but is prohibitively computationally expensive. Most importantly, the value of the AMS-DE is shown as the master process clearly balances the slave processes jobs very efficiently.

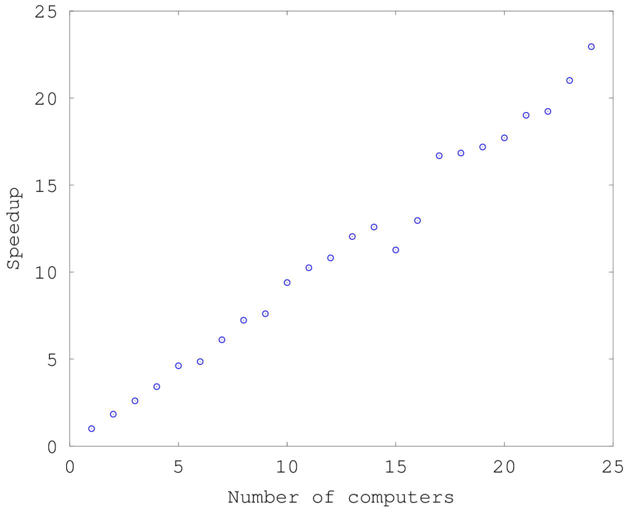


Figure 8: Speedup of the energy losses optimization on 1 to 24 computers. Although experiment was repeated 5 times on each number of computers, the speedup appears still very jagged, due to high variability in execution time of s a ingle cost function.

VI. Conclusions

In this paper, parallel performance of simulation-based optimization is demonstrated. On a small execution environment, composed of up to 24 computers, an almost linear speedup is achieved compared to execution on a single computer. Main finding is that a properly designed parallel evolutionary algorithm can use highly parallel systems very efficiently. Even if the simulator function is computationally demanding and does not execute in constant time. It has to be emphasized though, that the convergence of solutions towards the optimum slows down with increasing parallelism, but not significantly for low number of used processors. In this work, the slow down was not noticed, but mainly due to low number of performed repetitions.

The execution time plays a crucial role in realistic optimization processes. Bigger populations, as well as more detailed simulations are required for successful optimization process. The execution time increases dramatically when those two parameters are refined. The presented parallel implementation of simulation based optimization is capable of scaling well on modern parallel computers and could be employed for more complex optimization cases.

References

[1] 1: K. Price, R. M. Storn and J. A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization, 2005

[2] 2: A. E. Eiben and J. E. Smith, Introduction to Evolutionary Computing, 2003

[3] 3: B. Filipič and M. Depolli, Parallel evolutionary computation framework for single- and multiobjective optimization, 2009

[4] 4: D. W. Heermann and A. N. Burkitt, Parallel algorithms in computational science, 1991

[5] 5: I. Tomašić, A. Rashkovska, M. Depolli , R. Trobec, A comparison of hadoop tools for analyzing tabular data, 2013

[6] 6: M. J. Flynn, O. Mencer, V. Milutinović, G. Rakočević, P. Stenstrom, R. Trobec, M. Valero, Moving from petaflops to petadata, 2013

[7] 7: M. Depolli, R. Trobec, and B. Filipič, Asynchronous master-slave parallelization of differential evolution for multiobjective optimization, 2013

[8] 8: M. Depolli, V. Avbelj and R. Trobec, Computer-Simulated Alternative Modes of U-Wave Genesis, 2008

[9] 9: R. Trobec, M. Depolli and V. Avbelj, ECG Simulation with Improved Model of Cell Action Potentials, 2009

[10] 10: G. Kosec, Simulation of multiphase thermo-fluid phenomena by a local meshless numerical approach. , 2012

[11] 11: G. Kosec G, B. Šarler, Solution of thermo-fluid problems by collocation with local pressure correction, 2008

Session 4

Computer Networks, Distributed Systems, GRID and Cloud Computing

Overview of Software Defined Networking

Natasha Shipka
National Bank of the Republic of Macedonia
Skopje, Macedonia
naumovskan@nbrm.mk

Igor Mishkovski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
igor.mishkovski@finki.ukim.mk

***Abstract*— The main concept of Software Defined Networks is separation of the control plane (the way network devices forward traffic) from the data plane (network device that serves to forward traffic according to the control plane policy). Separated control plane is located in the central controller, which makes it possible to implement unique and flexible policies of forwarding traffic, limited only by software characteristics. Programmability (using OpenFlow as a communications protocol) is another important aspect of SDN in improving and simplifying network management. In this paper we are focusing on underlining the advantages of SDN and describing the characteristics of different types of SDN controllers (both commercial and open source). Thus, we present a comparative analysis of NOX/POX, Beacon, VMware (vCloud/vSphere), Nicira (NVP), Juniper/Contrail. In addition, the advantages and challenges of SDN, as well as areas of its implementation are further elaborated.**

***Keywords*— SDN; OpenFlow; comparative analysis; overview**

I. Introduction

What is Software defined networking? Software defined networking (SDN) is a type of network architecture that separates the network data plane (the network devices that forward traffic) from the control plane (the software logic that controls how traffic is being forwarded through the network). This decoupling of network's data plane and control plane allows the network operator to control the network behavior from single high-level control program. Software defined networking is changing the way communication networks are maintained, managed and secured [1].

Deployments of SDN are often used to solve a variety of network management problems in real networks. In current, traditional network architectures, network devices are bundled with a specialized control plane that has various features. This bundling effectively slows innovation. In conventional networks, once a flow management is defined, changing the configuration of the devices is the only way to make adjustment to the forwarding policy.

Software defined networking breaks these pieces apart. At the lowest level there are switches (devices that only forward traffic). On top of that there are more complicated control programs in the control plane that communicate with data plane through a well-defined software interface such as OpenFlow. On top of that control plane there might be more complicated applications written in higher level languages that perform management operations like traffic load balancing or security applications. This means that the control is moved outside of the network nodes to a centralized SDN controller.

II. SDN Architecture

SDN infrastructure consists of two parts: a control plane and a data plane. The control plane is the network's brain which is controlling the behavior of a network. It can run separately from the underlying network devices like routers and switches that actually do traffic forwarding. The control plane also computes the logic of how traffic will be forwarded through the network according to high level polices. It has the logic that controls the forwarding behavior in a network, for example routing protocols or network middle-box configurations [2].

The second part of SDN infrastructure is the data plane, which is typically programmable hardware that is controlled by the control plane. The data plane forwards traffic according to control plane logic, for example: IP forwarding, layer two switching, etc. Data plane is sometimes implemented in hardware. However, it is increasingly implemented in software routers as well.

There are many advantages of this kind of separate control: enabling more rapid innovations because control logic is not directly tied to hardware; enabling the controller to potentially see a network-wide view. Having a separate control channel makes it possible to have a separate software controller which facilitates the introduction of new services to the network much more easily.

Why separate the control and data plane? One reason is that by separating them, each can evolve and be developed independently. In other words software control of the network can evolve independently of the hardware. It means that one can buy routers, switches, middle-boxes etc., deploy them in the network and not be bound by the capabilities of the software that comes with the hardware at that particular time. The second reason to separate the control and data planes is that it allows the network to be controlled from high-level software program. High-level programs can control the behavior of the entire network and that way it becomes easier to reason about network behavior and it is also easier to debug and check this behavior.

There are various opportunities where separation helps: in data centers, enterprise networks, research networks, home networks and more.

Not only control plane should be programmable, but also data plane must be programmable as well.

SDN has three open interfaces: northbound (interface between applications and network infrastructure), southbound (interface between control plane and data plane) and east-west interface (interface between controllers)[8].

OpenFlow is a southbound API technology that provides control over switches. OpenFlow makes networks programmable. The controller can proactively or reactively add, update or delete flow over the southbound API.

The controller speaks to OpenFlow switch over a secure channel and the protocol effectively instructs the OpenFlow switch to update its flow table entries and to take different actions on various traffic flows that pass through a switch. The purpose of the control channel is to update this flow table. The logic concerning how flow table entries are updated is contained in the controller. The OpenFlow protocol specification defines number of things including components of the switch (flow tables – match and action, secure channels, group tables), the message format and what types of actions the flow table should be able to perform.

OpenFlow – capable network switches are available on the market from vendors, like IBM, HP, NEC, NetGear, etc.

Northbound API is programming interface (on SDN controller platform) that allows applications to program the network at a higher level of abstraction. Developers might use northbound API to implement more complex applications like firewalls, load balancer etc. The controller is responsible for insuring that those applications interact correctly. There are various uses of northbound API including path computation, loop avoidance, performing routing and security.

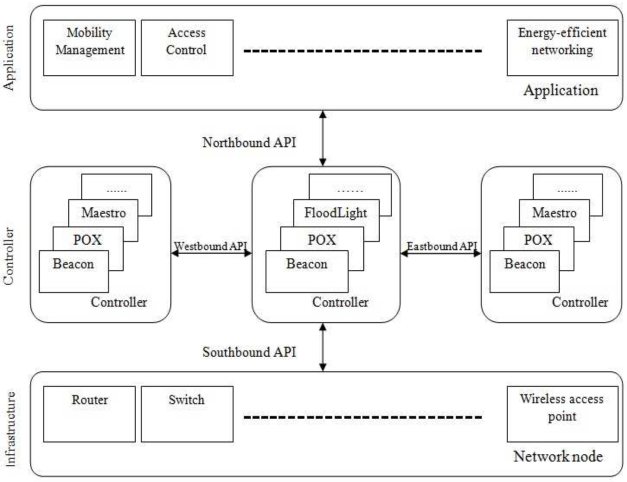


Fig. 1. SDN Architecture

Network virtualization should not be confused with software defined networking. Network virtualization is abstraction of the physical network that allows support for multiple logical networks running on a common shared physical substrate. It is a common set of physical routers, links etc. that support multiple logical network topologies on top of that physical infrastructure. The difference between network virtualization and software defined networking is that SDN does not inherently abstract the details of the physical network. It essentially separates the data plane from the control plane. Network virtualization is the technology that provides the abstraction so that multiple logical networks can be extended on top of that physical network. SDN can be a useful tool for implementing virtual networks.

III. COMPARISON OF SOLUTIONS

Recently there has been a vast commercial interest on software defined networking that led to many OpenFlow based control plane solutions.

Different controllers, either from commercial vendors or open source have different goals. Some are very focused on supporting production networks, whereas others are more focused and supporting education and research networks.

NOX is a first generation OpenFlow controller developed by Nicira and donated to the research community. Since year 2008, NOX is open source; it is stable and widely used. There are two versions of NOX: NOX-Classic which is written in C++ and Python, but it is not longer supported and the “new” NOX, which is C++ only, fast, has clean codebase and it is well maintained and supported. The programming model is very similar to programming models of many OpenFlow controllers, in which a controller registers for events and a programmer then writes event handlers that take specific actions or perform various tasks when those events are raised. NOX is bare bones implementation of the southbound API and does not provide many high level abstractions.

POX is essentially NOX in Python. It has several advantages over NOX: it has Pythonic OpenFlow interface; it has reusable sample components for topology discovery and path selection; it targets Linux, Mac OS and Windows. POX is widely used, maintained and supported, it is relatively easy to read and write code. The disadvantage is performance because it is implemented in Python. POX is very useful tool for rapid prototyping and experimentation, and hence it is very useful for research, experimentation and demonstrations, as well as for learning SDN concepts.

Beacon is open source, Java based, OpenFlow controller. It is fast, modular, stable, dynamic and supports both event-based and threaded operations.

Maestro is a centralized controller that uses multithreading for achieving better performance and scalability.

Floodlight is another open source, Java based, OpenFlow controller. It is developed by an open community of developers and is the core of a commercial controller product from Big Switch Networks. Floodlight is modular and the controller can be easily extended and enhanced.

In recent study [3] some controllers were evaluated on their ability to handle a large amount of control traffic. As shown in Fig.2, Beacon had the best throughput (over NOX, Maestro and Floodlight) because of its static partitioning and packet batching technique. In contrast, NOX faced synchronization overhead and therefore showed a reduced performance. The same study proved that Beacon had the best performance on switch scalability, (shown in Fig. 3), because of its static switch partitioning, packet batching technique and the low synchronization overhead.

Latency performance is the ability a controller to process the incoming packet-ins as fast as possible so it can quickly reply to the switches. This feature was also measured among NOX, Beacon, Maestro and Floodlight. Maestro had the best latency performance because of its workload adaptive batching technique that dynamically changes the batch size (Fig. 4). On the other hand, in the mentioned study, NOX, Beacon and Floodlight showed degradation in their performance as compared to Maestro because of their static packet batching design.

VMware provides a data center orchestration with its own SDN controller and agent implementation. The most famous VMware applications are vSphere, vCloud, vCenter and vFabric. The core of VMware solution is Java oriented.

Nicira's network virtualization platform (NVP) was released in 2011; it is more of a classic network controller meaning that network is the resourced managed. To compute, storage and image management, NVP works together with other cloud virtualization services. NVP works with open vSwith, which is the hypervisor soft switch controlled by NVP Controller cluster.

The Nicira controller provides a variety of northbound programmable APIs. It has network orchestration function and it is allowing a user to create a network overlay and link it to other management elements from vCenter/vCloud.

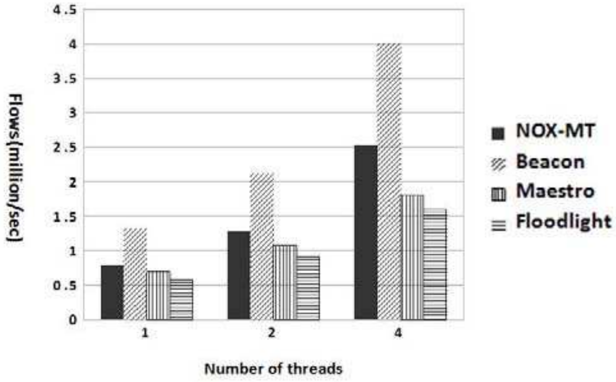


Fig. 2 Throughput performance [3]

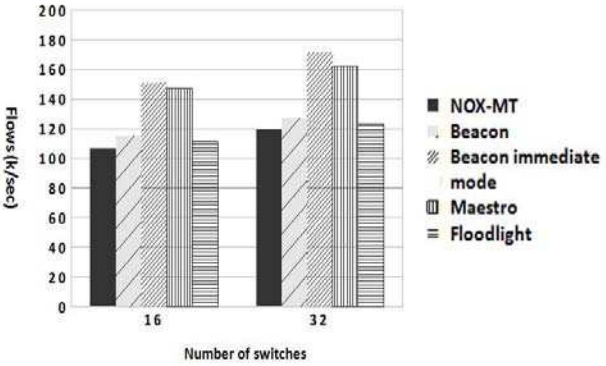


Fig. 3 Scalability [3]

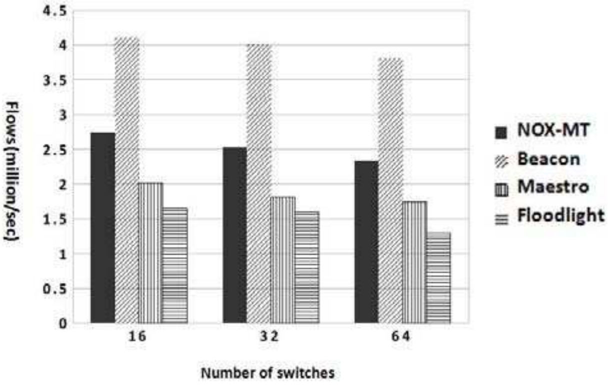


Fig. 4 Pre-packet latencies [3]

Juniper's Contrail Controller runs on top of Linux. The platform implements northbound API that applications and orchestrators can program to, including the OpenStack API integration. The main southbound protocol that Juniper uses is XMPP. However, additional southbound protocols such as BGP are implemented as well. The software for Juniper's Contrail Controller is contributed to OpenDaylight community which is a community devoted to maintaining open source implementations of various SDN control architectures.

Another SDN control architecture is Cisco's Open Network Environment which specifies a centralized software controller, a programmable data plane and the ability to provide virtual overlays.

Briefly, the SDN controller is the personification of SDN framework and is a reflection of the framework. When choosing the appropriate SDN controller, commonly there are ten criteria used to evaluate SDN controller: OpenFlow support, network virtualization, network functionality, scalability, performance, network programmability, reliability, security of the network, centralized management and visualization and the SDN controller vendor [10].

A. OpenFlow Support

Even though a vendor names a controller - "SDN controller" it does not mean it supports OpenFlow. That is why it is important to know whether the controller supports OpenFlow, which version of OpenFlow is implemented, as well as vendor's plan for implementing newer versions of OpenFlow.

B. Network Virtualization

Besides supporting the network virtualization that is well known and has been in production networks over a long period of time (like virtual LANs or Virtual Routing and Forwarding instances), SDN controller should also enable creating flexible, policy-based virtual networks.

C. Network Functionality

SDN controller should be able to discover various paths from source to destination and to split the network traffic across different links. Moreover, SDN controller is evaluated by the ability to define quality of service parameters on a flow-by-flow basis.

D. *Scalability*

When choosing an appropriate SDN controller, it is essential to know how many switches it can support

E. *Performance*

Significant criteria for evaluating SDN controller is its performance. Performance of the controller is measured by the time it takes to respond to data path requests and by the number of data path requests handled per second [9]. When choosing a controller, we must ensure that it is not a bottleneck in the entire network.

F. *Network Programmabiity*

As discussed previously, all SDN controllers have programmatic interfaces. The parameter "network programmability" of SDN controller refers to the ability to redirect inbound traffic and the ability to apply sophisticated filters to network packets.

G. *Reliability*

Reliability of a controller is a key factor for a reliable network. SDN controller must provide multiple network paths from source to destination. Hardware and software characteristics of the controller should be redundant, and controller ought to have the capability to be clustered.

H. *Security of the Network*

SDN controller has to provide an ability to apply class authentication and authorization to provide better security of the network. It needs the ability to limit the rate of the control communications and to notify if a malicious attack occurs.

I. *Centralized Management and Visualization*

One of the key benefits of software defined networking is centralized management and visualization of the network. The controller should provide end to end traffic flow visibility and be able to present a visualization of both, the physical and the virtual networks. In addition, SDN controller should have the possibility to be monitored using standard protocols and techniques, for example using SNMP.

J. *The SDN Controller Vendor*

The vendor of the SDN controller should be able to support the development associated with software defined networking and keep up with the rapid development of the SDN environment.

IV. SDN IN ACTION

There are various domains in which SDN applications can make network management easier, such as data centers, wide-area backbone networks, enterprise networks, internet exchange points, home networks.

In data centers, it is relatively common to move a virtual machine from one physical location to another as traffic demands shift. For example, Yahoo is using more than 400000 virtual machines, which is 1024 distinct inter host links between any pair of virtual machines for the topology Yahoo deployed. In this case it is problematic to guarantee a sub-second migration and to guarantee consistency in a network as that migration occurs. The solution that Yahoo has taken is to program each of the network switches from a central data base. As virtual machines migrate, the central controller knows about the migration and can update the switch state accordingly so that network paths update in accordance with the virtual machine migration [2].

Moreover, in data centers, separation of control and forwarding plane can significantly lower the cost of running a data center. Looking at a typical data center with 200,000 servers and a fan-out of 20, the requirements to support this data center topology is about 10,000 switches. If we take a particular switch from a standard vendor that costs 5000 US dollars, we are talking about 50 million USD just to deploy the switches. On the other hand, if we could deploy commodity switches that cost about 1000 USD, now the switch deployment costs about 10 million USD. So, for a large service provider that has 10 data centers, like Yahoo, Google, Facebook, the savings can reach 400 million dollars, and this amount could be used to hire engineers to develop control systems for controlling those commodity switches. The benefits of that separate control result in bigger flexibility, ability to tailor network for specific services and ability to quickly improve and innovate [2].

Separation of planes makes it easier to manage a data center through more flexible control in the context of addressing.

There are applications of software defined networking that solve problems in real world data center networks. One of those applications for data center is cloud computing. In cloud computing environments clients can take advantage of the fact that a network of numerous servers that is shared among multiple tenants can be scaled dynamically for that client as demand fluctuates. Data centers have typically multiple tenants, or independent users, thus allowing the operator of the data center to amortize the cost of the shared infrastructure.

There are different service models for cloud computing: software as a service model (the provider licenses applications to users as a service), platform as a service model (the provider offers software platform for building applications), infrastructure as a service model (the provider offers raw computing, storage and network).

SDN helps to solve many problems traditional data centers have. The PortLand design introduces a Fabric Manager which is like an SDN controller. This logically centralized fabric manager in combination with positional pseudo MAC addresses allows a layer 2 network to scale to thousands of servers inside a data center. In PortLand architecture each host has pseudo MAC address. The hierarchical structure of the pseudo MAC addresses allows the switches to forward traffic simply based on the structure of the MAC address. Fabric manager solves MAC learning and resolution of ARP requests.

One example for future possible use of software defined networking in wide area networking is in internet service providers (ISPs). Today's business relationships between ISPs are still extremely monolithic. SDN allows potential for ISPs to design and implement much richer and much more flexible set of interconnection policies. Time will tell whether ISPs will take advantage of this increasing flexibility and should SDN be deployed on a much broader scale in the wide area.

An interesting case study of software defined networking is B4 – a deployment of SDN in Google's wide area backbone network. Google operates two large backbone networks, one that caries user traffic and the other backbone network caries internal traffic between data centers. Managing both backbone networks is difficult. SDN, in particular OpenFlow has helped Google to improve backbone network performance and to reduce the complexity and cost of running a backbone network. Google runs many WAN intensive applications and the problem is that the cost per bit does not always decrease as the size of a network increases. The solution that Google has chosen to deploy is so called WAN Fabrics, which manages the WAN as a fabric, not as a collection of individual boxes. Unfortunately, current equipment and protocols make this particularly challenging because many protocols are box-centric, they provide little support for monitoring, low latency routing, fast failover and other types of mechanisms that are needed to provide this fabric abstraction. In contrast, a centralized traffic engineering approach is used in Google, where a centralized controller allocates the paths between each pair of endpoints to satisfy the capacity constraints of the network. The advantages of centralized traffic engineering are better network utilization since the controller has a global picture of the topology, faster converges to target optimum set of routes when a link fails, more control and the ability to specify intent, the ability to mirror production event streams for testing and the ability to use modern server hardware for the controller which results in better performance [2].

The deployment of software defined networking in wide area network can also help testing. In a decentralized network operators must do a full scale replica of a testbed to test new traffic engineering features. On the contrary, centralized control can use real production network as an input to research and test new ideas and features.

V. Advantages, Disadvantages and Challenges

There are numerous advantages of software defined networks over conventional network architecture. SDN simplifies the coordination of a behavior among a network of devices (for example to load balance a traffic across network or to make sure the security polices don't interfere with load balance polices). These types of coordination are much easier to handle in software defined networks than in conventional network where the network operator would have to configure and control each device independently in a low level vendor specific language.

Software defined networks are much easier to evolve. Conventional networks have vertical integration, so it's very difficult to evolve because the software is bundled with the hardware.

Controlling the behavior of the network from a high level program potentially makes the network behavior easy to reason about. It is much easier to look at a single program and to figure what that program is going to do and how it's going to control the network.

SDN makes it easier to apply conventional computer science approaches (programming languages, software engineering, testing) to old network management problems. These techniques cannot be applied in conventional networks because of low level operating.

Other benefits of implementing SDN are: dynamic access control, seamless mobility or migration, server load-balancing, network virtualization, using multiple wireless access points, energy-efficient networking, adaptive traffic monitoring, and denial-of-service attack detection.

From all the published literature we came across about software defined networking, disadvantages of SDN were almost not mentioned. So instead of disadvantages, we would refer to these as challenges.

The first challenge of separating control and data plane is scalability. Once the control elements are separated from the forward elements it is likely that one control element may be responsible for many forwarding elements (sometimes thousands). The other challenge is reliability or security: What happens when a controller fails or is compromised? We should hope that the forwarding elements continue forwarding traffic business as usual, but once we have separated the brains of the network from the devices that are actually forwarding the traffic, correct behavior is no longer guaranteed if a controller has failed or is compromised.

One solution for the security challenge is transport layer security with mutual authentication between the controllers and switches. Furthermore, SDN architecture supports network forensics (threat identification and management), security policy alteration and security service insertion (applications like firewalls and intrusion detection systems can be inserted) [6].

SDN's security is as good as the defined security policy. Authentication and authorization mechanisms can be implemented to ensure better security.

Another SDN challenge is interoperability or the way SDN solutions can be integrated into existing networks. One thing is certain, in order to achieve a hybrid SDN infrastructure (in which conventional and SDN enabled network nodes can operate harmoniously), further developments are required.

It is important to remember that software defined networking is just a tool. SDN does not specify a particular application nor is a solution. To take full advantage of software defined networking, the compelling application that internet service providers and operators want and need must be identified. Only then operators, ISPs and others will start pressuring vendors to include various SDN features in their network devices.

Martin Casado, the father of SDN and one of the founders of Necira, in an interview with Enterprise Networking Planet discussed the origins and the future of software defined networking. Even though Casado is one of the creators of SDN revolution in Stanford, he was no longer certain what the term SDN really meant in the market today. "I actually don't know what SDN means anymore, to be honest," Casado said. He marked that the term was coined in 2009, and at that time, SDN did mean something quite specific. "Now it is just being used as a general term for networking, like all networking is SDN. SDN is now just an umbrella term for cool stuff in networking. The goal is, how you make networking have the properties of software systems as far as innovation, provisioning speed, and upgrade speed," Casado said. "You want networks to be as flexible and as agile as compute is. That's not the case today, but that's where we're going" [11].

VI. Conclusion

In summary, there are many new frontiers for software defined networking. One of the most exciting challenges for SDN is to be recognized as much broader perspective than only OpenFlow protocols and match/action primitives. SDN is about separating control plane and data plane and using control plane to perform orchestration that may otherwise be extremely difficult.

Software defined networking has its focus on the following key features:

- Decoupling control and data plane,
- Centralized controller and centralized view of the network,
- The devices on the control plane are connected with open interface with the devices in the data plane,
- Network programmability by external applications.

The motivation of SDN is accelerating innovation in existing networks so that technology can be introduced more rapidly.

SDN develops networks built on standard hardware, so with a programmable interface built on a general-purpose hardware, a multivendor equipped network becomes a possibility. It promises an efficient and adaptive network with centralized control, flexibility and open interfaces between nodes.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the University “Ss. Cyril and Methodius”.

References

[1] T. Nadeau, K. Gray, “SDN:Software defined networks”, O'Reilly Media, September, 2013.

[2] N. Feamster, “Software defined networks”, Course on Coursera.org, 2013.

[3] S.A. Shah, J. Faiz, M. Farooq, A. Shafi, S.A. Mehdi, “An architectual evaluation of SDN contollers”, IEEE ICC 2013, pp.3504-3508.

[4] H. Kim and N. Feamster, “Improving network management with software defined networking”, IEEE communications magazine, pp. 114-119, February 2013

[5] L. Vanbever, J. Reich, T. Benson, N. Foster, J. Rexford, “HotSwap: Correct and Efficient Controller Upgrades for Software-Defined Networks”, HotSDN '13, 2013.

[6] S. Sezer, S. Scott-Hayward, P.K. Chouhan, B. Fraser, D. Lake, J. Finnegan, N. Viljoel, M. Miller, N. Rao, “Are we ready for SDN? Implementation challenges for software defined networks”, IEEE communications magazine, pp. 36-43, July 2013.

[7] S. Agarwal, M. Kodialam, T.V. Lakshman, “Traffic engineering in software defined networks”, 2013 Proceedings IEEE INFOCOM, pp.2212-2219, April 2013.

[8] M.K. Shin, K.H. Nam, H.J. Kim “Software-defined networking (SDN): A reference architecture and open APIs”, 2012 International Conference on ICT Convergence, pp. 360-361, October 2012.

[9] A. Tootoonchian, S. Gorbunov, Y. Ganjali, M. Casado, R. Sherwood, “On controller performacne in software defined networks”, Hot-ICE 2012, April 2012.

[10] Ashton, Metzler & Associates, “Ten things to look for in an SDN Controller”, 2013.

[11] OpenFlow inventor Martin Casado on SDN, VMware, and SDN hype, http://www.enterprisenetworkingplanet.com/netsp/openflow-inventor-martin-casado-sdn-vmware-software-defined-networking-video.html, April 2013.

Centralized System for Cloud-based Mobile Services for Students

Kostadin Mishev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
kostadin.mishev@finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*—By spreading of the biggest computer network Internet recent years, cloud services have become the common used method for publishing the public information to the end-users [1]. Also, mobile phones, especially smart phones, brought access to that global network closer to the user due their portability. So, implementing services will allow better interaction between the providers of information and end users. Their communication would be facilitated and improved.**

Faculties are one of the institutions that should have reliable and on-time delivered information to students. Faculty of Computer Science and Engineering is one of them. We built a system for students whose goal is generating a channel for on-time and reliable information delivery between students and their professors. It is allowed by mTomen system which aims to bond these services with mobile application and allow better interaction in information exchange. It uses CAS for user central authentication, so all services are available by one login.

Currently, mToken system is used for authentication, evidence of students' presence on lectures and sending push notifications from other registered services to the user. Also, it has open interfaces for scalability with other future services. In this paper, the system architecture and the possibilities of scalability of mToken would be described in details.

Keywords—cloud-based mobile services; Google Cloud Messaging; internal geolocation; presence evidence

I. Introduction

Today, 98% from the youth in age from 18 to 29 use a mobile phone [2]. 83% are using smart phone and 63% of youth have 3G Internet access. In 54 countries, the universities have open access to Internet for all students allowed by using the Eduroam service. Eduroam (**edu**cation **roam**ing) is the secure, world-wide roaming access service developed for the international research and education community [3]. Faculty of Computer Science and Engineering allows access to this service by setting access points all over its campus. So, students have free access to Internet simply by logging in using their existing CAS account.

All of these numbers give the opportunity of implementing online services accessed by mobile phones. These cloud-based services for mobile applications should be centralized. Centralization for this kind of services means:

a) One database for evidence of students using the mobile application

b) Single sign-on protocol for allowing access to any service

c) Channel of communication between mobile application and service

d) One service for sending push notifications

e) Scalability for future services

f) Determination of internal geographic location– providing personalized services by the geographic position of the user in the campus

In implementing these features for mToken system, there were obtained several issues. As first, the CAS implementation does not provide login by proxy service for token validation. The CAS is unable to validate user logged by a mobile application. Second, there should be implemented a mobile application that could communicate with these services using a secured channel. So all services provided by mToken, should be based on HTTPS (SSL) connection for data exchange. Afterwards, there should be implemented a system with artificial intelligence to decide the right position of the user in the faculty campus. This should be the cheapest solution based on the current infrastructures. Next issue is creating a universal system for sending push notifications from all services to any user (student or professor). Every user should be identified by unique id that references his mobile device (registration id) and it should be saved in database for future use (Figure 1). In next sections, all of these issues would be explained in details and specified appropriate solution.

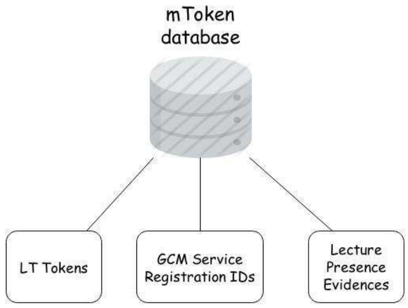


Figure 1. System components

II. Related Work

As we mentioned above, the need of on-time informing the students has a big significance especially in academic environment. There are few companies that have started software projects related with real-time informing the students and their parents with news about the school, lectures and the curriculum. They use SMS Text, Text-to-Voice, Voice-to-Voice and Email like communication channels. Also, they have strict set of services that they have allowed to be used using their notification system. Such services are: Rediker notification system [4], SchoolMessenger [5] , K-12 Alert [6], uNotifiy e2Campus [7] etc.

III. System Architecture

A. CAS Mobile Authentication

The prime goal of mToken system is enabling one login authentication for all services provided for mobile application. It is based on CAS authentication, so the same usernames and passwords are in use. Main issue obtained by implementing this service was making validation of tokens obtained from CAS login [8] (Figure 2). The session of accessing the login page by the mobile application, must be transferred to the service, in this case, module that makes the validation of the obtained token. That validation proves the identity of the user that owns the token [9]. After the successful validation, mToken system generates long term token which is sent back to the mobile phone and could be used for future access to the services without CAS login (Figure 3). Long term token have own time to live which currently is set up on 3 days. So, on the third day, the user should login again with its CAS username and password. After the successful login, he could access all other third-party student services of FCSE that are allowed for validation from mToken system, only by using the long term token (Figure 4).

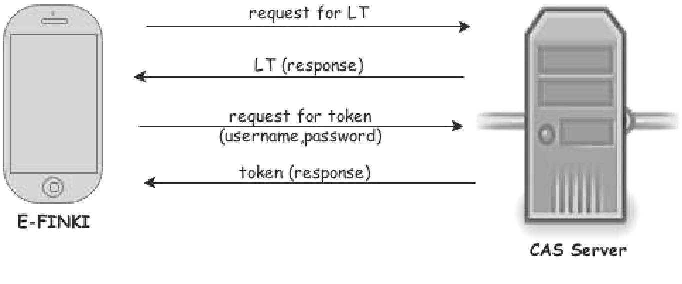


Figure 2. E-FINKI requests CAS server for CAS token

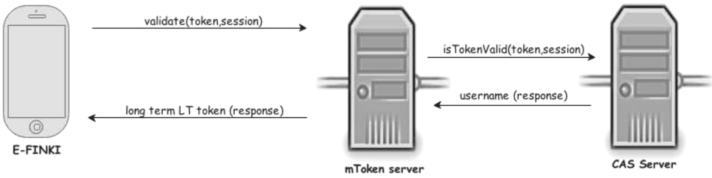


Figure 3. Generating the long term mToken token

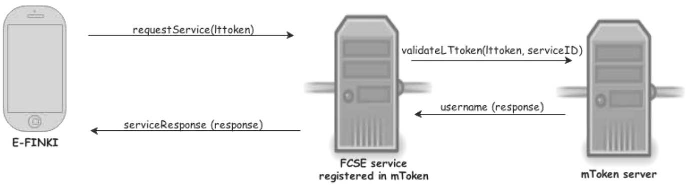


Figure 4. Authentication of user to third-party student services

One of the advantages of mToken service for generating long term tokens is that it does not make any exchange with the user's username and password. The mobile phone application accesses the login page of CAS system, so it only receives the token which is the result of validation between mobile phone and CAS system. The service should only sustain the session of the user until it validates the token and obtains the identity of the user that received that token. This means that this system is secured and does not abuse the privacy of the CAS user.

Creating permission for validation of long term token from a service, could be made easily from the administrator of mToken system by simply inserting the host name (URL) of service in the mToken database.

B. Centralized Module for Sending GCM Messages to Students

Google Cloud Message (GCM) is popular method for on-time presenting of new information to the users of mobile phone applications [10]. It enables sending messages to users of Android and iOS applications as well [11]. Using the mToken system, it could be sent different types of messages. The content depends from the purpose of the service and it is individual. GCM messages are generic and any messages following specified format, could be handled by E-FINKI mobile application. Only services validated from mToken could send GCM messages to students.

Every user of the E-FINKI mobile application obtains registration id when he installs the application from the market. That registration id is saved on file in the mobile memory card [10]. When the user logs in in the mToken system, the registration id is sent to the mToken database and it is associated with his username. One username could have multiple registration ids (Figure 5).

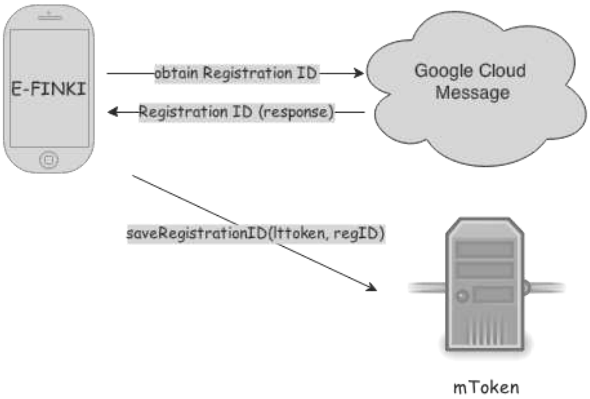

E-FINKI
obtain Registration ID
Registration ID (response)
Google Cloud Message
saveRegistrationID(lttoken, regID)
mToken

Figure 5. Registration of mobile device to GCM

Afterwards, when the service registered to mToken, wants to send any message to the user, it only sends as POST parameter the content of the message and the id of his student index. The mToken system, remaps the index with registration id obtained from his phone, and sends the message to him. The student obtains the message as push notification on his mobile device on time (Figure 6). The URL where the service should send the message is following:

https://mtoken.finki.ukim.mk/services/push/android/invokepushtostudents

The POST parameters should have the following format:

message: *String* (the message to be send to user)

indices: *JSONArray* (array of ids of student indices)

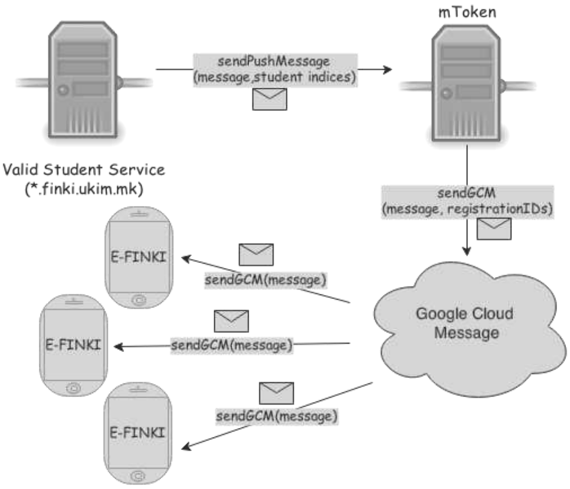


Figure 6. Process of sending GCM message

This service, as part of mToken system, enables centralized and simple sending of GCM messages to students from any valid service. All registration ids are saved in one database and all kind of messages are handled from one application.

C. Centralized Module for Student Presence Evidence

mToken implements own module for storing records of students presence on lectures. It is opened module, so the evidence could be obtained by different kinds of systems for presence detection. All of proved data is saved in the mToken database. Afterwards, analyses of students presence could be obtained for any variable (course, professor, student) in the FCSE. Also, the professor could check all present students on his lecture using professor module on E-FINKI mobile application.

Currently, mToken integrates an own-build module called Wiloc for determination of internal geographic location using the wireless signals in the nearby. Wiloc implements algorithm which creates decision trees that are built by taking training set with BSSIDs and their signals strengths from access points near the classrooms of the FCSE. It is based on C4.5 algorithm used to generate decision trees [12]. Training set is taken from appropriate people by using the administration unit of E-FINKI mobile application. Access to that unit is allowed only to the staff of FCSE. Also, it gives an opportunity of rebuilding the model of decision trees by taking new samples for improvement in decisions and adding new rooms to the system. This recreation is simple and lasts only 30 seconds obtaining enough big training set from the location of the user (Figure 7).

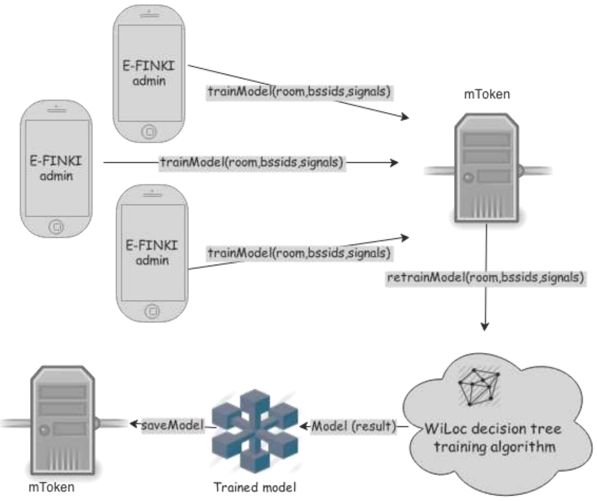

E-FINKI admin
trainModel(room,bssids,signals)
mToken
E-FINKI admin
trainModel(room,bssids,signals)
E-FINKI admin
trainModel(room,bssids,signals)
retrainModel(room,bssids,signals)
saveModel
Model (result)
WiLoc decision tree training algorithm
mToken
Trained model

Figure 7. Retraining of the Wiloc decision tree

The students could check-in to prove their presence in the classroom by using the student unit from the application E-FINKI. At first, Wi-Fi scanners in their mobile devices scan for the wireless signals and strength in their environment. Afterwards, the results are posted to the Wiloc module which determinates their location. If determined location is equal with the classroom specified in the schedule, the presence of the student is saved as record in the mToken database (Figure 8). The lecturer could review the checked students as a list anytime by using E-FINKI application.

Currently, the decision tree is obtained by 10 000 training instances in 8 rooms in FCSE campus and its precision is 91% in determination the correct location of the user.

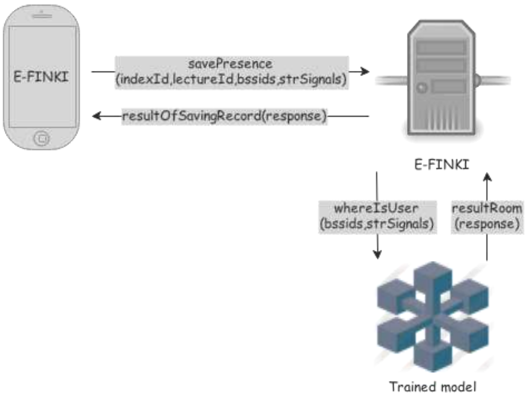


Figure 8. Creating a record of student presence

IV. Conclusion

In this paper was briefly described the system architecture of mToken with its currently functional services. Principal purpose of this system is offering mobile cloud-based services intended for students at Faculty of Computer Science and Engineering. It gives primary set of methods for manipulating the request made by mobile devices. Also, the channel of communication between the mobile device and mToken server is secured with HTTPS protocol, so it gives additional safety in information exchange with related services.

This centralized hub of student services differs from the other ones by its scalability. New services could easily allow parental control of the student presence. It permits simple authentication with CAS system differing from the others which have own system for identification.

Architecture of mToken allows easy scalability of the system with new mobile based services. Registering the service URL in the database as valid service, permits interaction with other mobile-based service and provides obtaining simple communication among professor-students and students-students using GCM messages.

The model of presence evidence module is made to be easy configured to work with other presence detection methods. Currently, Wiloc method is scalable because adding new room to the model lasts only 30 seconds. So, after 30 seconds every user of the system could check-in the presence in the room. It is cheapest, fastest and most efficient solution for indoor geographic location.

References

[1] Demarest G., Wang R.: Oracle Cloud Computing, Oracle, 2010

[2] The use of mobile phones between youth, www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet/

[3] Eduroam official website, https://www.eduroam.org/

[4] Rediker School Notification system, www.rediker.com/school-notification-system

[5] SchoolMessenger, http://schoolmessenger.com/solutions/about-the-network/

[6] K-12 School Notification Systems, http://www.alertsolutions.com/education/k-12-school-notification-system/

[7] e2campus uNotify, http://www.e2campus.com/higher-education/routine-notifications

[8] Official web site of JASIG CAS, http://www.jasig.org/cas

[9] CAS Specification, https://wiki.jasig.org/display/CASUM/Home

[10] Developping GCM mobile applications, developer.android.com/google/gcm/gcm.html

[11] GCM services and iOS services, https://cloud.google.com/developers/articles/ios-push-notifications

[12] Martinez J., Fuentes O.: Using C4.5 as variable selection criterion in classification tasks, IASTED-Spain, 2005

Example application of in-memory grid technology on a health system to reduce response time

Despot Jakimovski
Software Engineering
Faculty of Computer Science and Engineering, Ss Cyril and Methodius University (student)
Skopje, Macedonia
despot.jakimovski@gmail.com

Vladimir Trajkovik
Software Engineering
Faculty of Computer Science and Engineering, Ss Cyril and Methodius University
Skopje, Macedonia
vladimir.trajkovik@finki.ukim.mk

Recently, there has been a great deal of interest in big data through the use of nosql databases and in memory data grid technologies. The common deployment of these technologies is in case when application starts showing bottlenecks while handling large amount of data and application scalability cannot be performed by a traditional node addition.

Through the use of technologies that can divide the processing task to different units in parallel, the study will try to reduce the time of response in the example healthcare system.

The research present in this paper focuses on the use of big data in healthcare system that provides recommendation for users depending on the user profile determined by extensive manipulation of the big data obtained from various inputs (sensors, social networks, external information). We applied the mapreduce algorithm and produced two hundred samples of response time that was sublimated in average response time. The lowest average response time was achieved with the in-memory data grid technology.

The paper also gives the correlation between our research and the different applications of big data in healthcare.

Keywords— map reduce; big data; cloud; personal healthcare system; social network;

I. INTRODUCTION

Recently, there has been a great deal of interest in big data through the use of nosql databases and in memory data grid technologies. Their application is wide.

Medicine and informatics have been together for almost half a century [1]. As the informatics progresses, the use of informatics in medicine grows bigger. The extent of that is shown in a collaborative healthcare system that gives recommendations and suggestions for preventive intervention instead of the use of medical emergency and hospital admission. Apart from the use of medical data of the user for assessing which activity made the user's health better [6], the system is used in daily activities for recommending future activities, through current states, in real time [2] [7].

The amount of data for one such healthcare system is being measured in millions of records and the memory requirements are expected to fluctuate with the growth of the number of users. That implies the need for a scalable healthcare system. By using tire based architecture, scaling the application (adding a node or processing unit to the system) will not better the performances linearly [8] [9] [10]. This is solved by using a space based architecture offered by in-memory data / compute grid technologies. Such collaborative healthcare system has also the requirement for a smaller round time during comparison of data in the current activity. The in-memory data / compute grid technologies help with speeding up the response by: working with the primary memory (RAM) instead of the secondary memory; processing the task in parallel by dividing it among processors in multiple partitions / processing units, lowering the time significantly [11] [12] (this also gives the added benefit of using commodity hardware instead of expensive multiprocessor servers); and saving data in a database (if needed) asynchronously (hence no time lost in synchronization).

This type of technologies is applied in many systems. SETI institute uses a form of grid computing called CPU-scavenging [3]; distributed document system where comparison is made between two technologies of this type Sphere and Hadoop [4]; bioinformatics applications [5]; European Union supports grid computing through its development funds.

World Health Organization with the Disease Outbreak News, even in 1996 was gathering information with the aim to predict and inform on disease outbreak which is in turn a big data problem. The continuation of that medical area was further extended by globalincidentmap.com, outbreaks section and HealthMap in 2006 and Google Flu, launched 2008.

As early as 1998 there is an effort and vision for the personal electronic health records to become centralized and shared in the amount that the patient, laws and health institutions allow [20]. This trend took up in Switzerland from 2008 (VitaClic, one of the companies, that provide keeping personal health records), by incorporating big data in the health sector, improved “the quality and efficiency of healthcare and lower its costs"[13]. There have also been legislative activities to stimulate the big data need in medical history. An example of that is the American Recovery and Reinvestment Act (2009) that was intended to “encourage physicians to adopt electronic medical record-keeping systems” where “physicians and providers share health information across jurisdiction" [14]. By having more effective record keeping, physicians and health providers could analyze the data for more information, in turn revealing the need for big data. Not only these type of applications will make health record keeping [18][19] more centralized and more widely available for the customer/patient, but will "contribute to formulating research priorities, researching the causes and epidemiology of disease, assessing the effectiveness of preventive interventions and clinical care, paying providers based on their performance" [17], "acquire more reliable and timely information about the cost and quality of health care" [14]. In 2008 there was a pilot done using an ATM-like system for medical record keeping for medically-underserved populations [17]. The system that we want to speed up is a continuation of the same intent to "anonymize and graph and predict health trends across populations." [17].

Another appearance of big data in healthcare is linked to the patients need to take care of their own health by researching online, exchanging experiences on social networks, using more than 2000 Yahoo groups (tied to health), searching Google [14].

Big data plays or will play a vital role in medical services such as post-marketing drug surveillance, providing information about patient populations, their health conditions, and the specific drugs that people are using, getting custom medical care based on genetic analyses [14], coordination between multiple specialists and higher risk patients [15], change people's poor behavior pattern [16].

Cloud, and with that Big Data technologies, are maturing their place in the technology scene [18].

As wireless sensor network technologies can be used for providing wide range of healthcare services for people with various degrees of cognitive and physical disabilities, delivering high-quality care services for babies and little children of families where both parents have to work, control of home appliances, medical data access, and emergency communication for residents and their caregivers, constant medical monitoring [21], and quality of data sampling and delivery (trustworthiness) [22], in this paper we give an extension of the usability of these networks for giving recommendations and suggestions for preventive intervention through collaborative algorithms.

Apart from the challenges of output transmission power of sensor nodes, real-time availability and reliable data measurement and communication issues, security, energy efficiency and consumption, privacy of pervasive healthcare systems [21][18][2], we focus the research on the scalability, reliability and availability through investigating the responsiveness of the data analysis. These characteristics of the application are important and for some systems might need to breach even some privacy requirements [22].

The research done on the COHESY system [2] [23] gives an explanation on the way the data is gathered ("bionetwork, data for the user's physical activities (get by the mobile application), medical records of the user (obtained from clinical centers) and data of the user profile on social network (so far based on knowledge of social network)"), the way it is sampled, and provides an example recommendation algorithm "for users to carry out physical activities to improve their health". The research present in this paper focuses on finding the appropriate cluster for a user profile by researching suitability of mapreduce technologies by comparing two of them in order to find a better candidate for executing the recommendation algorithm.

In the next section we present the methodology used in our research. Then we present obtained results and performance analyzes. Last section concludes the paper.

II. TECHNOLOGIES

A. Gigaspaces

Gigaspaces is an in memory data grid technology that allows scalability, low latency and high-throughput by implementing the JavaSpaces specification that in turn is based on the Tuple Space concept (shares information between tuples instead of sending messages).Domain objects (entries) are stored by the space acting as in-memory service. The specific actions that can be taken against the entry are write, read, take (read plus delete from the space), and notify (when the entry registers changes). A different combination of these actions can derive different paradigms like caching by using write and read since entries are stored within the application; parallel processing by using write and take since the master/worker pattern (originally invented within the context of Tuple Space) can be achieved multiple writers providing entries and consumers processing them; messaging by using write/take and notify since asynchronous informing occurs.

A group of multiple space instances form a cluster, which is viewed by the user as a one big space. The way they are organized is called a clustering topology. There are two main clustering topologies (replicated and partitioned) and a hybrid. The most widely used is the hybrid resilient (from its replicated part) partitioning clustering topology or Partitioned-Sync2Backup. It gives the ability for the data to be both partitioned when the data doesn't fit in one space instance, and replicated to handle a failing space instance.

The mapreduce ability is supported by the openspaces remoting feature which is based on the spring remoting which in turn is based on the java remote method invocation (RMI). There are two implementations of remoting supported by gigaspaces: Event Driven Remoting that is made available by the Polling Event Container that implements the pooling consumer pattern and is good for queue semantics and master worker pattern, but doesn't support mapreduce; and Executor Based Remoting which allows for map/reduce invocations in a synchronous or asynchronous manner that also can failover transparently (if a primary space instance fails, it will transparently route the remote invocation to the backup). The Executor Based Remoting uses Task Executors. They are made available by OpenSpaces and they are collocated with the space instance. So they execute a task from the client which sends the task through a proxy to a processing unit (space instance) through remoting. Executes the task there and returns the result through the proxy again to the client.

The client implements the Task interface (define what the task will execute on the space instance) and sends it through remoting on the specific space instance / processing unit by providing the routing id or sends it to multiple processing units (space instances) or the whole cluster. By sending to multiple processing units, it can also wait on all the results from the processed tasks on the remote space instances, and reduce them all together, in effect doing the mapreduce pattern. Gigaspaces provides a mechanism (DistributedTask execution) so the client doesn't have to do this manually. This is called distributed task execution or broadcast task execution (if all the primary space instances from the cluster are involved).

B. MongoDB

MongoDB is a NoSQL technology which also allows scalability and replication. Its structure consists of collection namespace and index namespace (contained within extends organized in doubly linked lists). Indexes are placed in memory and collections taken into memory on demand.

MongoDB solves scaling with sharding which scales the application by adding additional nodes - shards without adding complexity to the application. Three types of sharding are available. We are using hash-based sharding that generates MD5 hash, based on the document shard key, so these documents are distributed uniformly across the shards. Range-based and custom tag-aware sharding exist.

The sharded cluster consists of shards (one master and one or more replica slaves), config servers hold the number and locations of shards and data within them (counting 3), and mongos instances or query routers that route queries from the client to the appropriate primary/master shard (referring to the config servers to find out where the appropriate shard for their query is).

When data / documents are requested, they are placed in memory. Updates on documents are done in memory by marking it dirty and thousands of operations are done before the systemdecides to write the changes to disk. This makes the processing rather fast. Since documents are padded (space left for additional data to be written), when writing dirty documents to disk, the changes either overwrite a previous field, or add some data which can use up the pad. The document will need to be rewritten to another address only when the pad is filled. The padding size is also decided dynamically by keeping record how fast fields within a document change, and with this minimize disk I/O. There are other optimizations that MongoDB takes to handle disk memory client communication and to insure better throughput. One way is keeping the documents in chunks with equal size, which in turn are placed in the shards. There is a balancer which redistributes chunks to a shard least congested, as they grow out of size. It supports many types of indexes unique, compound (on multiple fields), geospatial (for location queries), sparse (for indexing fields only in case when they exist in the document) and text search type. We are using a sparse unique index. Furthermore, there is a query optimizer that takes turns in executing different indexes of the same queries from time to time when the collection hasn't changed, and chooses the faster index. Also if a query asks for data that is only contained in an index, MongoDB doesn't have to look on disk (it returns the data from the indexes in memory). Another optimization is the mechanism in which MongoDB is syncing the changes to disk. The process is journaling and instead of writing the changes to the collection on disk (it might have to wait for other operations to finish and takes more time), it writes the deltas (differences) on disk in a separate journal file and later when allowed writes these changes to the appropriate collection. The tradeoff for this architecture is the inability to provide ACID transactions on multi-document level, which must be handled within the client application.

Before map phase starts, the documents are taken from the collection into memory (if not already there) and any sorting and filtering can be applied (and is recommended since MongoDB won't have to write the intermediate sorted/filtered results to disk). A map javascript function maps the results from the documents by emitting key-value pairs. If multiple values appear for a key than the reduce phase kicks in and reduces the results with the reduce javascript function. Additionally a finalize stage can be issued if we need to do some extra calculations after the reduce stage is finished. Mapreduce results can be written in a collection (and that further can be used for some subsequent mapreduce operations on the same growing data set - incrementally) or results can be inlined, but the collection has to be less than 16MB. The map function's emit can hold maximum 8MB. The emit function can be called zero, one or multiple times. The reduce function can be invoked on the same key multiple times (the result of the reduce becomes the input parameter for the subsequent reduce on the same key). Mongodb can take, as expected, sharded collections. When the collection is sharded the mapreduce is executed on each shard in parallel.

C. Differences

The main difference between these two technologies, relevant to our research, is the place of operation execution. The operations are memory to memory in the case of Gigaspaces and memory and disk (file based) for MongoDB. Since both of the technologies have concept of keeping not yet replicated or saved changes on file (MongoDB through journaling, Gigaspaces through the redo log), this is not taken into consideration although they might have slight differences in the way they do the change logging.

III. METHODOLOGY

The model used was created from the SportyPal application. The calculated values from a workout consist roughly of the following parameters: start and end time, start and end gps latitude and longitude, distance, total time, calories burned, best two thousand meters, maximum speed and pace, average pace and heart rate, total climb and descent. These parameters were used to create calculated minimum, maximum, average and total from all the workouts calculated values that count forty eight, from here on, grouped calculated values. These were used to compare the users (user profiles) and to check which cluster they belong to. The cluster represents a logical concept that can be assigned to a group of user profiles that have same parameters or parameters that diverge slightly. For the purpose of this research we consider the case where clusters will not be used for direct comparison, so their state in our case is not shown. Taken this into account, the clusters contain the ids.The ids of the clusters the user belongs to are collocated with (replicated in) the user profiles. This enhances the speed since there would be no need to search the entity in another partition (space instance in Gigaspaces, shard in MongoDB).

The experiment starts by executing the jar on command line and providing initial number of clusters and users that need to be created, whether or not we want the initial feeder ran, how many times should the comparing be done, the number of partitions that the comparison would be ran on, and the boundary percentage of clusters that the user will belong to (or the cluster subset percentage), all as command line parameters. Since the number of clusters for a given profile will be limited, when generating the profiles we set the maximum of the clusters that one profile could belong to is three percent of the total. Since the testing was done for hundred clusters in total, initial users could belong to a maximum of three clusters. The number of users was set to five million. Partitions were set to eight (and sixteen in the case of gigaspaces 16 partitions), and number of repeating comparisons set to two hundred.

Firstly the generator creates the clusters within the application, and writes them to the appropriate partition/shard using the routing id. Each cluster has a random routing id assigned to it between and inclusive one and the number of partitions (eight). After the clusters are saved in partition, the users are assigned a random number of clusters they belong to between 0 and three, and they too are routed to one of the eight partitions.

After the initial feeding has been finished, the comparison iterations start. It consists of creating a user that doesn’t belong to any clusters which is going to be used to be compared to the users in the partitions. After the user is created, the measurement of time is initiated. Then a task (CompareUserDistributedTask for Gigaspaces and mapreduce mongo operation for MongoDB) is sent and invoked on all partitions. Within that task the map function compares thirteen of the forty eight grouped calculated values of the user that needs to know which clusters it belongs to, and the one from the partition. The number of grouped calculated values was chosen randomly for the purpose of this research. This map compare phase consists of filtering out all those profiles that achieved a matching percentage lower than a matching percentage boundary (decided by the cluster). The matching percentage refers to the number of matching grouped calculated value pairs met. In our case, a grouped calculated value pair is matched if they are equal (although, we could compare values in any way, for instance we could measure the deviation of the values depending on a cluster’s value deviation boundary). And the matching percentage boundary was set to zero percent (in order to get as many possible candidates). The approach presented in this paper is inspired by the methodology presented for the COHESY system [2] in terms of the need to calculate the value (membership) of the profile in relation to certain cluster. Once the comparison was done, the clusters the partition user belonged to, are added in a map (if not already there) for the ‘new’ user and their counter incremented. Also a total count of users for that cluster is saved. In the reduce phase a sum of the total cluster users and a sum of the matched cluster users is created. And the final calculation mandates that the ‘new’ user will belong to the cluster if and only if the percentage of matched users from the total of the users belonging to the clusters is below some limit (which in our case was set to ten). Once the calculation is finished the time is stopped and logged, and another comparison iteration is started. In the case of eight threads used with MongoDB, eight of the iterations were done in parallel.

The logs contain the amount of time it took to populate the system with the initial users and clusters, the amount of time it took compare the first user to the ones in the partitions (since for certain technologies like MongoDB, the data that is not in memory is taken from disk the first time the application wants to read something, although this could be seen from the other iterations), and the amount of time it took for the two hundred subsequent compares.

IV. RESULTS AND PERFORMANCE ANALYZES

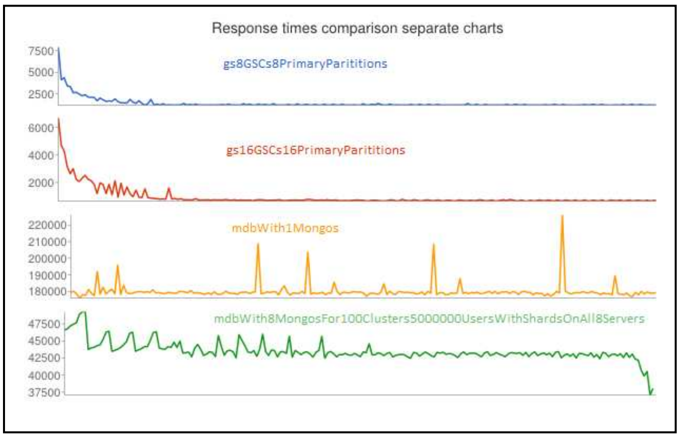


Fig. 1. Response times comparison in separate charts

Gigaspaces with eight primary partitions drops exponentially with local maximums bigger than average response time at the iterations 1, 3, 7, 10, 15, 18, 20, 25, 28, 32 although only the iterations 1, 3 are above the local maximum average of 2853.2 ms. With 16 partitions also drops exponentially but it has eleven local maximums bigger than the response time at 1, 6, 10, 15, 18, 20, 22, 24, 27, 30, 38 with iterations 1, 6, 10 being above the local maximum average of 2406.64 ms. The graph normalizes after 32 and 38 iterations. Whereas with MongoDB in the case of one used router when sending one mapreduce task at a given time, the response times stick to the average with occasional scattered peaks. The local maximums above the response time average are at iterations 2, 7, 10, 12, 15, 17, 19, 29, 42, 54, 65, 74, 82, 91, 108, 110, 123, 125, 134, 156, 169, 187 and local maximums above the local maximum average (189038.59 ms) being 10, 17, 65, 82, 125, 169, 187. In the fourth experiment MongoDB runs as many possible mapreduce tasks in parallel as available processor cores (eight in our case). The local maximums above average response time are 7, 16, 24, 28, 32, 36, 38, 40, 46, 53, 57, 60, 65, 68, 71, 77, 79, 88, 104 and local maximums above the local maximum average (45214.84 ms) being 7, 16, 24, 32, 53, 60, 68, 77, 88. So for MongoDB, the normalization doesn't occur at all or happens after the 104th iteration. For the later MongoDB scenario, since, at the beginning, the documents/collections are still not in memory, the response times are considerably higher. There is a slight increase each iteration since some of the threads will always return later, even if their mapreduce tasks were sent at approximately the same time. This happens because they compete for the same shared resources and one will always have a higher priority than the other. As time passes the processor core 'attacks' become distributed uniformly, since the main application threads send the mapreduce in a more uniform approach. At the end, it takes less time to process a mapreduce task because there are less mapreduce tasks for the processor to process, hence they are not competing for the same shared resources. Similar uniformity is expected when a single mapreduce task is split and multiple threads are used. The threaded MongoDB approach works 4.164 times better than the non threaded since MongoDB has no division of the mapreduce task within the processor. It processes the mapreduce task in one core.

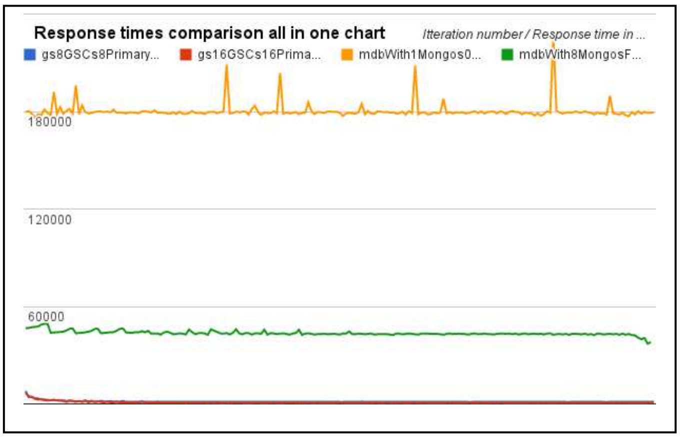


Fig. 2. Response times comparison all in one chart

Like for MongoDB, in Gigaspaces using more partitions per server (number of partitions approaching number of cores), will make better utilization of the resources, since it too takes better advantage of the processor cores. We could have divided the mapreduce task the same number of times as partitions multiplied. It should have rendered a similar result where we would see a fall in response time. The average response time when using sixteen partitions against eight gave 34.35% better response time. As in the case with MongoDB the more utilization of the cores we make, the more expressed the peaks will be.

On figure 2 we can see the significant difference in response times between the two MongoDB approaches. The first approach where one mapreduce task is sent at one time on the 8 servers (orange/upper one), has a response time average of 180.238 seconds. Whereas, the second approach, where eight mapreduce tasks are sent to all eight servers (green line) achieve 43.517 seconds round time. Compared to the Gigaspaces 1.377 seconds and 0.904 seconds average response time, MongoDB lags behind mainly because of the need to transfer the files from disk to memory.

TABLE I. TABLE AVERAGES AND DISTRIBUTION STATISTICS

	Gigaspaces 8 partitions	*Gigaspaces 16 partitions*	*MongoDB 1 mapreduce task per 8 servers*	*MongoDB 8 mapreduce tasks per 8 servers*
Avg. (sec.)	1.377	0.904	180.241	43.284
Std. Dev.	0.439	0.569	5.097	3.379
Sample Variance	0.192	0.324	25.979	11.419
Population Variance	0.191	0.322	25.849	11.362
Population Std. Dev.	0.437	0.567	5.084	3.371

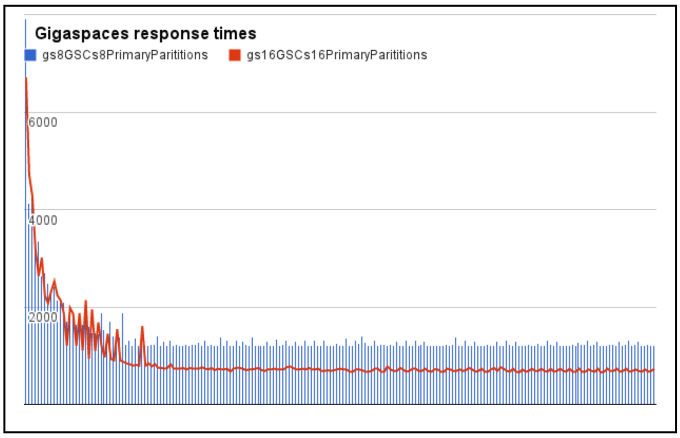


Fig. 3. Gigaspaces response times comparison

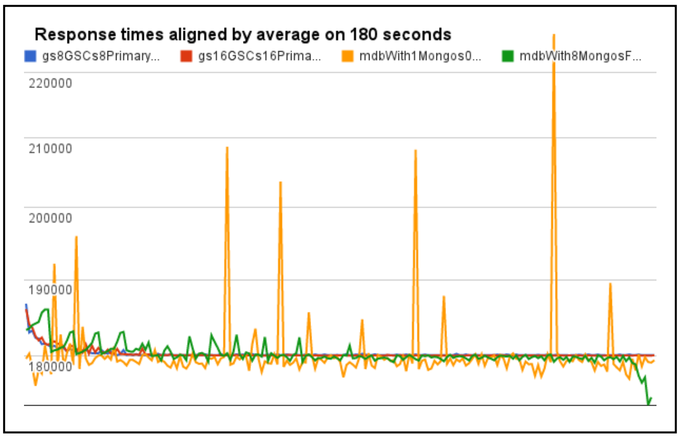


Fig. 4. Response times aligned by average on 180 seconds.

In our case, taken the above, using Gigaspaces will provide better scheduling of tasks.

V. CONCLUSION

By using Gigaspaces, an in memory data grid technology, and MongoDB, a NoSQL technology we inspected the low boundary of the response time for a collaborative healthcare system represented by the model of the SportyPal application.

Using in memory data grid technologies will render the best results. Specifically, by using two partitions for servers with eight cores. The response time gets better when the number of partitions approaches the number of CPU cores, but since this was not linear (34.35% increase when changing from one partition on server to two partitions on server), it remains to be tested how much it could change with more than two partitions.

As in the related work done on bioinformatics paired data processing [5], where they compared using Mapreduce technologies (from multiple used), we come to the conclusion that Mapreduce technologies are a preferred approach when classifying, because of their capability to hold large data sets, and to collocate the computation with the data. Therefore, they simplify solutions over traditional systems.

A further research can be done by exploring some options in both technologies.

MongoDB mapreduce can be speeded up by using the splitVector command to split the collection into multiple parts and use multiple threads to process the parts. Currently MongoDB doesn't split the collection automatically so that it can make use of the multiple cores on a processor like in Hadoop. This is why presplitting might render better results. Instead of splitting the collection for one mapreduce task, we've shown the same, by running eight mapreduce tasks in parallel. The presplit approach needs to also work with different databases, since each mapreduce task needs to write the output and the same lock will need to be acquired if one database is used. By changing the size of chunks kept in a shard response times could also be tested. Since our map phase of the task is not sortable, we can't improve on the speed by using a sort on the collection prior to issuing the mapreduce.

Adding partitions could be tested for speed and we could see what will the correlation be when the number of partitions approaches the number of cores.

Additionally, other NoSQL (like HBase and Hadoop) and other data grid technologies could be compared.

REFERENCES

[1] "The History of Health Informatics" Health Informatics, Nursing Informatics and Health Information Management Degrees, University of Illinois at Chicago, 2011

[2] Vlahu-Gjorgievska, E., Trajkovik., V. "Towards Collaborative Health Care System, Model – COHESY", 3th IEEE Workshop on Interdisciplinary Research on Ehealth, Services and Systems (IEEE IREHSS 2011). IEEE Press, New York, 2011

[3] "SETI@Home Credit overview". BOINC. Retrieved April 21, 2010

[4] Yunhong Gu, Robert Grossman, "Toward Efficient and Simplified Distributed, Data Intensive Computing", IEEE Transactions on parallel and distributed systems, January 2, 2010

[5] Jaliya Ekanayake, Thilina Gunarathne, and Judy Qiu, "Cloud Technologies for Bioinformatics Applications", IEEE Transactions on parallel and distributed systems, January 2011

[6] E. Jovanov, "A Wireless Body Area Network of Intelligent Motion Sensors for Computer Assisted Physical Rehabilitation," J. NeuroEng. Rehab., vol. 2, no. 6, 2005

[7] Jamil Y. Khan, Mehmet R. Yuce, F. Karami, "Performance Evaluation of a Wireless Body Area Sensor Network for Remote Patient Monitoring", Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, Vancouver, BC, 2008

[8] Nati Shalom, "Space-Based Architecture and The End of Tier-based Computing", White Paper, GigaSpaces Technologies Ltd, 2006

[9] Nati Shalom, "The Scalability Revolution: From Dead End to Open Road" An SBA Concept Paper, GigaSpaces Technologies Ltd, 2007

[10] Gigaspaces, "Scale Up vs. Scale Out", White Paper, June 2011

[11] Jaliya Ekanayake, Geoffrey Fox, "High Performance Parallel Computing with Clouds and Cloud Technologies", Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Volume 34, 2010

[12] Lavanya Ramakrishnan, Keith R. Jackson, Shane Canon, Shreyas Cholia, John Shalf, "Defining future platform requirements for e-Science clouds", SoCC '10 Proceedings of the 1st ACM symposium on Cloud computing, ACM New York, NY, USA, 2010

[13] Olivier Brian, Thomas Brunschwiler, Heinz Dill, Hanspeter Christ, Babak Falsafi, Markus Fischer, Stella Gatziu Grivas, Claudio Giovanoli, Roger Eric Gisi, Reto Gutmann, Matthias Kaiserswerth, Marco Kündig, Simon Leinen, Willy Müller, David Oesch, Marius Redli, Didier Rey, Reinhard Riedl, Andy Schär, Andreas Spichiger, Ursula Widmer, Anne Wiggins, Markus Zollinger, "Cloud computing", White paper, Swiss Academy of Engineering Sciences, 2012-11-06

[14] David Bollier, "The Promise and Peril of Big Data", The Aspen Institute, Communications and Society Program, Washington, DC, 2010

[15] Richard Hillestad, James Bigelow, Anthony Bower, Federico Girosi, Robin Meili, Richard Scoville and Roger Taylor, "Can Electronic Medical Record Systems Transform Health Care? Potential Health Benefits, Savings, And Costs", HealthAfafirs, September 2005

[16] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell, Dartmouth College, "A Survey of Mobile Phone Sensing", IEEE Communications Magazine, September 2010

[17] Nathan Botts, MSIS, Brian Thoms, PhD, Aisha Noamani, Thomas A. Horan, PhD, "Cloud Computing Architectures for the Underserved: Pulbic Health Cyberinfrastructures through a Network of HealthATMs", Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010

[18] Minqi Zhou, Rong Zhang, Wei Xie, Weining Qian, Aoying Zhou, "Security and Privacy in Cloud Computing: A Survey", Sixth International Conference on Semantics, Knowledge and Grids, 2010

[19] Rui Zhang 1,2 and Ling Liu, "Security Models and Requirements for Healthcare Application Clouds", IEEE 3rd International Conference on Cloud Computing, 2010

[20] Ilias Iakovidis, "Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe", International Journal of Medical Informatics 52, 1998

[21] Hande Alemdar, Cem Ersoy, "Wireless sensor networks for healthcare: A survey", Computer Networks 54, 2010

[22] JeongGil Ko, Chenyang Lu, Mani B. Srivastava, John A. Stankovic, Andreas Terzis, and Matt Welsh, "Wireless Sensor Networks for Healthcare", Nov. 2010

[23] Kulev, Igor; Vlahu-Gjorgievska, Elena; Trajkovik, Vladimir; Koceski, Saso, "Development of a novel recommendation algorithm for collaborative health - care system model", Computer Science & Information Systems, Vol. 10 Issue 3, Jun. 2013.

Implementation of RIAK DB-Engine for optimising the storage and processing of images

Bane Georgievski, Veno Pachovski, Eva Blazhevska
School of Computer Science and Information Technology
University American College Skopje
Boulevard 3-ta Makedonska Brigada, Skopje 1000, Macedonia
bane.georgievski@gmail.com, {pachovski, blazevska}@uacs.edu.mk

***Abstract*—The use of graphics and images on the Web is increasingly prevalent, from everyday blogs and forums to serious and complex data analysis from the global economy, industry, science and many other areas. Hence, optimizing the storage and processing of images arises as a challenge. The required optimization can be achieved through the extraction and storage of the needed parameters of the image and applying the methods and techniques to reduce the network load while operating the system. The purpose of this paper is to investigate the feasibility and suitability for utilizing the unconventional RIAK database in the application for determining the similarity among images. The architecture and robust characteristics of RIAK promise good results when working with large data sets. In addition to improving the performance of the application, an algorithm is proposed for effectively determining the similarities among the images.**

Keywords—image analysis; data storage; search optimization; big-data; NoSql;

I. Introduction

The exponential growth of data has spawned an era of rapidly evolving database systems and architectures that fundamentally rethink the concepts and techniques conceived and developed to manage and access data. As web technologies advance, a growing multitude of prospective internet users store vast amounts of data on a daily basis making the inevitable need for a fast and reliable database engine. The last forty years, relational database management systems dominated the database market unrivaled, but this way of doing things has proven inadequate when working with large data sets. Recently, a new generation of data management systems have arisen, mostly non-relational, and are proving to be effective solutions to the needs of an increasing number of large-scale applications. They are classified as "NoSQL" databases. This database model is not suitable for every application, mainly due to the lack of full ACID transaction support. That is why it is common today to utilize hybrid systems, running both SQL and NoSQL database engines.

As humanity drowns in data, the means to filter and organize the content of the data, gets more sophisticated every day. This paper attempts to provide a starting ground for a solution to the issue of finding and identifying similarities between images.

Image processing is an expanded area, dedicated towards improving and optimizing image processing algorithms. [1] and [2] propose an efficient way of parallelizing execution of this algorithms, however the focus in these papers is set only on accelerating algorithms' execution.

Another approach recently popularized with the rise of specialized hardware technologies, is testing and evaluating new processor architectures, implemented on DSP or FPGA. [3] presents a collection of suggested architectures dedicated for image processing. Some of those organizations were taken into consideration and their realization lead to positive results [4], [5].

This papers' approach is completely in other direction. Its purpose is to develop an application for the issue of finding and identifying similarities between images, but for the first time implementing this type of system/application, inside a database system, in our particular case, Riak database engine.

The paper is organized in five sections. Section two presents a brief introduction and explanation of perceptual hash algorithm, its properties, advantages and reasons for choosing it for this paper's work. Furthermore, section three consists of a brief overview of Riak database engines and its best features, continuing with blending pHash and Riak in section four - the practical realization of our project. Afterwards, this paper continues with presentation of the results from the executed tests, analysis and discussion. The paper concludes with section 6, where the general conclusions are made.

II. Perceptual Hashing

Many studies in the field of cognitive science, psychology as well as natural sciences have conveyed that people are visual thinkers. As much as we'd hate to admit it, appearance is everything. Only by looking at something one can immediately decide if he likes it or not. This is because human beings are visual creatures. According to 3M Corporation and Zabisco, 90% of information transmitted to the brain is visual, and visuals are processed 60,000 times faster in the brain than text. Throughout the years, many researchers work and experiments evolved around this phenomenon and contributed with thorough studies substantiated/supported with massive amounts of data in form of analyses, media, papers etc. Furthermore, todays' cutting edge technology enables acceleration and optimization in processes and applications in every academic, industrial or even life aspect. Hence, a utilization of intensive computer processing was introduced in image analyses, quickly becoming a trend and attracting attention of many theoretical researchers and application developers. Image processing refers to processing of a two dimensional picture by a computer. Basic definition is: An image defined in the "real world" is considered to be a function of two real variables, for example, a(x,y) with a as the amplitude (e.g. brightness) of the image at the real coordinate position (x,y), [6], [7].

As mentioned before, modern digital technology has made it possible to manipulate multi-dimensional signals with systems that range from simple digital circuits to advanced parallel computers. The goal of this operation can be allocated in three groups: Image Processing (image in - > image out), Image Analysis (image in -> measurements out) and Image Understanding (image in -> high-level description out). The aim of the studies in this field is clear, and is for computers to adapt the way that human brain works and resonates images, so for this same aspect to be implemented and exploited in various application, contributing in better marketing, data mining, or in other words, understanding and connecting different aspects of the web and digital media genuinely.

All image processing techniques actually are mathematical functions of operations with arrays (matrixes) since images in computer science world can be presented digitally in different format. For these mathematical functions to be executed on a computer machine they need to be translated into algorithms. As a result from all the work done in this area, there are plenty proposed algorithms for image processing, however all of them differ in terms of performance, accuracy, efficiency, requirements and consequently implementation. As with everything else, one should choose the most appropriate one for his needs, according to an application. For the purposes of this paper, a fast algorithm was required, resistant to rotation, skew and color variation, thus Perceptual Hash algorithm was chosen.

A fingerprint of a multimedia file can be generated by extracting number of features from it, and represent them in some predefined format. Perceptual hashes have the ability to create similar hashes for similar images, contrary to the cryptographic hash functions where even though images have some level of similarity between them, their hashes may be completely different. Being robust enough to survive transformations or "attacks" for a certain input, and at the same time remaining flexible enough to recognize a difference between dissimilar files, is a must for perceptual hashes. These so called "attacks" may come in a form of rotation, skew, contrast adjustment and different compression/formats.

For these reasons perceptual hashing is considered as e engaging field of study, with a great potential for computer science research.

Different application can implement pHash algorithm, starting with copyright protection, similarity search for media files, or even digital forensics. The goal of this research/paper being a web image crawler, perceptual hash algorithms meets all the requirements. Figure 3, gives an example of input image to an application deploying pHash algorithm and the corresponding output in form of binary fingerprint.

This new image hash is based on Marr wavelet which calculates the perceptual hash based on edge information with particular emphasis on corners [8]. According to [9], it has been shown that the human visual system makes special use of certain retinal cells to distinguish corner-like stimuli. It is the belief that this corner information can be used to distinguish digital images that motivates this approach. Basically, the edge information attained from the wavelet is compressed into a fixed length hash of 72 bytes. Binary quantization allows for relatively fast hamming distance computation between hashes, [10], [11].

The idea of web image crawler wraps around given image going through an image analyze process which will extract certain features from it, and according to them will query database storage of prospective similar images, linking that particular image to the most related one. For this concept to be realized and effective there is a firm need of a huge knowledge base (in our case fingerprints of lots of images) to which an image can be compared to. Thus, next important requirement for this project is effective data storage, capable of storing massive amounts of data, at the same time not very complex for querying and suitable for scaling. Riak meet the requirements and was chosen as underlying platform for our web image crawler application.

III. Riak distributed data base system

Riak is one of many new non-relational database structures that work by distributing data across multiple servers. It is designed to deliver maximum data availability. One of the fundamental tradeoffs for this system is – high availability in exchange for possibly outdated information [12]. The most profound things about Riak are:

- High-volume: Vast amount of data stored in the database is available for read and write when needed
- High velocity: Easily upgradable, responds to growth very well.
- High-variety: Any type of file can be stored as a value within the database.

Even if a system requires storing and serving terabytes of data a day, it will respond with exceptional speeds for requests.

IV. RIAK COMPONENTS

Riak is a Key/Value (KV) database, particularly designed and developed to spread data across clusters consisting of physical servers (nodes). These database systems are built upon key/value architecture, or in other words, the basic idea of how Riak works, is that you can store a value with an immutable key and retrieve the data when needed.

Riak's fundamental principle revolves around buckets, keys and values. Values can be data of any type, and the only way to access/retrieve this value is trough bucket/key pair. Buckets provide a logical namespace for objects stored in the database. This means that objects with identical keys can be stored in different buckets without conflict. Furthermore, essentially important is the physical platform on which Riak relies on, which defines the way Riak gains its functionality and main attractive features, like high-availability and distribution. Riak cluster is what keeps this system together.

Moreover, the major concept which defines the organization of Riak cluster is known by the name Replication and Partition. In order to retain high availability at all time, Riak's default feature is to replicate data into multiple storages. Thus, the main idea is that if one storage node goes down, another one will serve the requests.

The second part from Riak's concept, addresses the problem of increasing capacity, through dividing data into multiple places so performances are gained only by cheap hardware investment. Thus, by implementing this method of replication and partition of data, Riak cluster is organized in a way appropriate for fast querying, at the same time enabling highly available storage. Central to any Riak cluster is a 160-bit integer space which is divided into equally-sized partitions.[13]

Each physical server represents a node from the cluster, and each node contains virtual nodes - vnodes inside. These virtual nodes are elements of a clusters' (sometimes referred to as ring) and occupy certain part of it. Their number can be determined by associating with the number of physical nodes in the cluster. The best practice is to use an example. If we have a ring including 32 partitions which should be divided into 4 nodes (servers), the result will be, 8 virtual nodes for each physical server. This example is given on Figure 1.

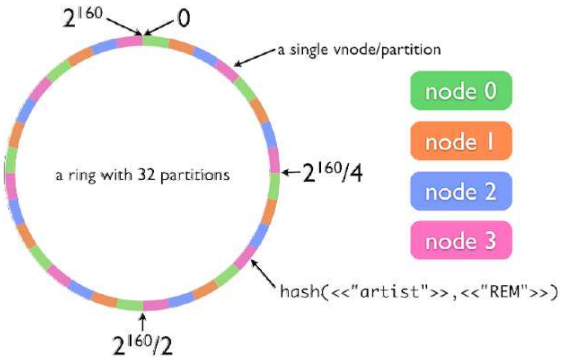


Fig. 1: Riak Ring

Continuing, worth mentioning is one of Riaks' best features for which it's known and valued, its Search capabilities. Riak has several built in features for querying and searching data.

1. A distributed full-text search engine with a robust query language, [12-LRB]
2. Tagging objects with additional values and query by exact match or range.
3. MapReduce implementation for sophisticated analyses for large datasets.

Regarding its features Riak database systems can be employed in variety of applications. Some use cases where retailers use Riak are:

- Shopping Carts: Using architectural principles from Amazon, Riak provides an always-on low-latency customer experience.
- Product Catalogs: Designed to scale out rapidly accommodating fast-growing volumes of images, product and user data.
- Mobile Applications: Fast and reliable small object storage to power highly concurrent mobile experiences across platforms.

The combination of simple database storage in key/value form and the capability for implementing powerful search made Riak an invincible choice for fulfilling our projects' goals. The next section will be dedicated on further explanation of how exactly Riak was implemented and configured to meet the requirements of a web image crawler.

V. Implementation

For the purposes of this paper, we installed Riak on a UNIX machine utilizing the PHP API provided by Basho. The images, no matter the size or format, need to be analyzed so that the necessary data can be extracted and stored in the database. It is also possible to store the whole image as a binary value in Riak but there are limitations to the size of the file to approximately 50MB. For now, the path to the image stored on the local machine will suffice.

The image analysis process consists of the following steps:

- Reducing the size - As Average Hash, pHash starts with a small image, larger than 8x8; 32x32. This step aims to simplify the DCT computation, not because it is needed to trim down the high frequencies.[13]
- Reducing the color – Process of reducing images' color to grayscale just to further simplify the number of needed computations.
- Computing the DCT. The DCT separates the image into a collection of frequencies and scalars.[14]
- Reducing the DCT - Keeping only the lowest frequencies in the picture.
- Computing the average value - As the Average Hash, computing the mean DCT value is required (using only the 8x8 DCT low-frequency values and excluding the first term since the DC coefficient can be significantly different from the other values and will throw off the average). This completely excludes flat image information like solid colors, from being included in the hash description.
- Further reduction of the DCT – This is the most important step. Setting the 64 hash bits to 0 or 1 depending on whether each of the 64 DCT values is above or below the average value. The result does not give the actual low frequencies; it just shows the very-rough relative scale of the frequencies to the calculated mean. The resulting bit array will not vary much as long as the overall structure of the image is

still the same, meaning it can survive gamma and color histogram adjustments without complications.

- Constructing the hash – Converting the 64 bits into a 64-bit integer. The order does not matter, just as long as it is consistent.

Once the image hash is created, it needs to be stored in the database along with the additional information. Following is presented Table with detailed information.

TABLE I: BACKEND STORAGE PERSPECTTIVE

Bucket		
Data Type	***Key***	***Value***
An image file in the format of .JPG .PNG, .GIF or other image format.	The pre-calculated hash value of the image	Basic information about the image: Header, Source

*The calculated hash of the image is stored as the key

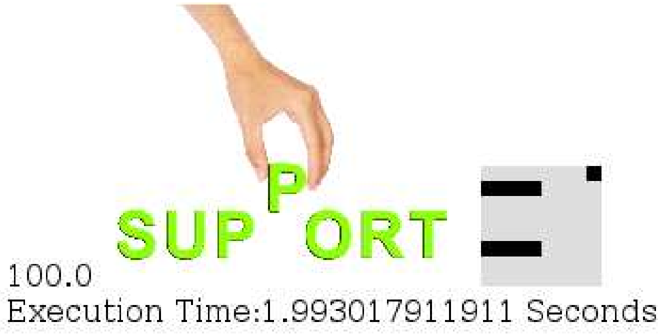
Following are presented the times for each phase of the application execution.

Time required for single image processing and constructing appropriate hash has been measured as 0.038638830184937 Sec, which is quite long time, compared to other hash algorithms. This is a consequence of the intensive processing needed for extracting images' features. Inserting hash as key, along with the image URL as a value takes about 0.0029208302497864 seconds. A conclusion can be made that image processing is a decisive bottleneck in this application and by accelerating this process, a large amount of time can be saved. This was the reason for further experimentation with different algorithms, more particularly dHash.

Testing dHash algorithm with the same image used in evaluating pHash performances, reviled the following results.

- For processing the image: 0.089251041412354 Seconds.
- For storing the image: 0.003881692886352 Seconds.

After comparing pHash time of execution to other popular hash algorithms, frequently used in image processing, it was revealed that perceptual hashing is the most effective and accurate for this type of application. Thus, this project proceeded with implementing pHash algorithm as the image processing component and Riak database as backend storage.



```
100.0
Execution Time:1.993017911911 Seconds
```

Fig. 2:Results generated when searching with pHash

Fig. 2 is a visual representation of the results generated when searching for an image in a bucket with 1000 different images. From left to right: the similarity percentage, the image we searched and a visual representation of the hash generated for the image.

Applications implementing Riak, or NoSQL data in particular, can gain additional acceleration by employing Map Reduce method. MapReduce is a method of aggregating large amounts of data by separating the processing into two phases, map and reduce, that themselves are executed in parts. Map will be executed per object to convert/extract some value, and then those mapped values will be reduced into some aggregate result. What do we gain from this structure? It's predicated on the idea that it's cheaper to move the algorithms to where the data lives, than to transfer massive amounts of data to a single server to run a calculation [12]. This principle can be a foundation for continuing this project by employing the database backend storage on several Riak nodes.

Next, the results of the project's testing are presented. Figure 3, presents the performance of this application, where x-axis gives the number of images present in the storage at the moment of executing the php script and y-axis gives the time execution in seconds.

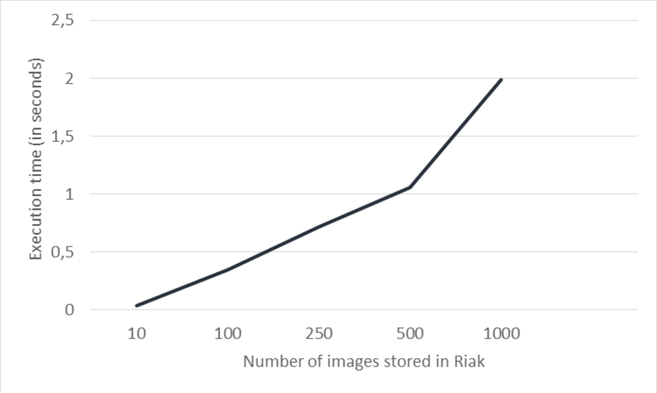


Fig. 3: Riak execution performances

Since execution time is a sensitive and important feature of a system, detailed results are given in Table II.

TABLE II: DETAILED EXECUTION TIME OF RIAK IMAGE SEARCH

Number of pictures	***Execution time***
10	0.035221099853516
100	0.34610676765442
250	0.71643304824829
500	1.0596771240234
1000	1.9901819229126

Even though detecting similar files typically involves complex heuristics and/or O(n^2) pair-wise comparisons, by using a hash function that hashed similar files to similar values, image similarity could be determined simply by comparing pre-sorted hash key values. This technique was combined with the powerful MapReduce method in Riak, and managed to reduce performance execution of our application to O(n) complexity, which can be noticed on Figure X, by keeping linearity of this application functions, despite the number of images stored in the database.

VI. CONCLUSION

The purpose of this paper was to demonstrate the efficiency of a NoSql Database (Riak) to store and process images. Because of the high minimum requirements of running a Riak database, we weren't able to see the full power of managing large amounts of data with this DB Engine, but we believe that further testing and appropriate hardware will reveal that RIAK is up to the task. Another task for the future will be to optimize the algorithm for detecting image similarities and obtain a O(log n) complexity performance.

REFERENCES

[1] Bruce Draper, Walid Najjar, Wim Böhm, Jeff Hammes, Bob Rinker, Charlie Ross, Monica Chawathe, José Bins, "Compiling and Optimizing Image Processing Algorithms for FPGA's", Department of Computer Science Colorado State University.

[2] Rinker, R., et al. "Compiling Image Processing Application to Reconfigurable Hardware", in IEEE International Conference on Application-specific Architectures and Processors. 2000. Boston

[3] Best papers from Design and Architectures for Signal and Image Processing 2007 & 2008 & 2009, "Algorithm-Architecture Matching for Signal and Image Processing".

[4] Sparsh Mittal, Saket Gupta, S.Dasgupta "FPGA:An Efficient And Promising Platform For Real-Time Image Processing Applications", Proceedings of the National Conference on Research and Development in Hardware & Systems (CSI-RDHS 2008), 2008.

[5] Tamás Raikovich, Béla Fehér, "Application of Partial Reconfiguration of FPGAs in Image Processing", 6th Conference on Ph.D. Research in Microelectronics&Electronics, 2010.

[6] Young, Ian Theodore, Gerbrands, Jan Jacob, Van Vliet, Lucas Jozef Delft University of Technology, "Fundamentals of Image Processing", ISBN 90–75691–01–7.

[7] Tinku Acharya and Ajoy K. Ray (2006). Image Processing - Principles and Applications", Wiley InterScience.

[8] D. Van De Ville , M. Unser, "Complex Wavelet Bases, Steerability, and the Marr-Like Pyramid, IEEE Transactions on Image Processing", v.17 n.11, p.2063-2080, November 2008.

[9] www.phash.org/Docs - Design and Develpment Guide

[10] Christoph Zauner, "Implementation and Benchmarking of Perceptual Image Hash Functions", University of Applied Sciences Hagenberg, 2010.

[11] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Werner Vogels , "Dynamo: Amazon's Highly Available Key-value Store", SOSP 2007 .

[12] Eric Redmond, "Little Riak Book".

[13] http://docs.basho.com/riak/latest/theory/concepts

[14] Andrew B. Watson, NASA Ames Research Center mage Compression Using the Discrete Cosine Transform, Mathematica Journal, 4(1), 1994, p. 81-88

[15] http://hackerfactor.com/blog/index.php?/archives/355-How-I-Met-Your-Mother-Through-Photoshop.html

Migrating Educational Services in the Cloud

Trends and Possibilities

Roza Arsova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, R. Macedonia
roza.arsova@gmail.com

Sonja Filiposka
ACSIC department
University of the Balearic Islands
Palma de Mallorca, Spain
sonja.filiposka@finki.ukim.mk

Abstract— **In this paper we give an overview of the cloud computing benefits for the educational institutions as well as risks that must not be overlooked when embarking on this type of venture. We review a number of success stories from all over the world about implementing cloud computing into the educational process on small and big scales. We also discuss the possibility of migrating the Macedonia's educational services into the cloud and the potential benefits and risks that this process contains. We present a possible integrated cloud solution design and discuss the steps needed in order to put the design into practice.**

Keywords—education; services; cloud computing; migration;

I. Introduction

The educational institutions all over the world are entering the cloud computing market being aware that the IT tools are becoming a necessity for accomplishment of the educational goals. Even more, the new age requires that schools are open towards the community so that its digital knowledge capacity is connected to the global school knowledge system. In other words, if, until recently, the schools were meant to provide capacity and ask for students to come in in order to provide them with knowledge, today's schools are supposed to put their knowledge out into the digital space, into the environment occupied by their students. Today's successful education is based on real time communication, as well as data and resource sharing.

Learning with computer support today has been translated from learning using a computer towards learning using the network as platform, that is from expert supported learning towards environment supported learning, from fixed data towards mixing and analyzing data. With this e-learning concept, cloud computing becomes the main pillar of the learning infrastructure.

With the rising popularity of cloud computing and its vast use of services via the commodity Internet, the IT companies have realized the interest in investment and development of free, or severely discounted, cloud service offers for the educational institutions. This trend is mainly due to the fact that they perceive the young students as the future customers of their commercial services. Also, by investing into a loyalty strategy, they reinforce their market brand. On the other hand, the educational institutions strive towards the cloud computing options in order to reduce their costs and increase effectiveness [1] while keeping lock step with new technologies that improve the learning process.

There are numerous success stories concerning the usage of some or a full range of educational services in the cloud in different educational institutions in the world. Although the implemented solutions are different, the main gains of the projects are equal: enhanced learning capabilities and cost effectiveness. Following the current world trends combined with the obvious benefits from cloud service are raising the question of implementing and taking advantage of the educational services in the cloud with the end goal being to migrate Macedonia's lower level educational system in the cloud.

Towards this aim, in this paper we give an overview of existing succesfull cloud migration stories together with a concept of a possible migration of the educational services for Macedonia that will ultimately lead to a modern, high quality cost effective education.

II. Cloud Education Examples

The cloud services for the education offered by the big names like Microsoft with its Live@Edu [2] project and Google with the Google Apps for Education [3] mainly include mail, calendar, office packets that include text editors, spreadsheets, presentations, graphics and web site editing and alike. The offer for this self-managed software that has no special hardware requirements is very tempting for the always-low educational budget. Another important cloud service that is of extreme importance for the educational institutions is the cloud data storage possibility with controlled access. This is especially the case for audio and video contents that are used as multimedia support in today's educational process. This type of content has no privacy issues since it is made publicly available with the teacher's control only on the content and not access, while it can benefit greatly from the cloud computing approach since it is meant to be accessed by a huge number of simultaneous users. Another obvious cloud migration candidate is the Learning Management System (LMS) that offers a vast range of support for the students and teaching staff. However, in this case, the full set of data is not public and great care must be taken in order to secure the private data and ensure disaster recovery and backup options especially in the case when the complete student record will be stored somewhere in the cloud.

Some of the grand names inclined on using Google Apps for Education include the University of Notre Dame, Georgetown University, Monash University, Sapienza University of Rome, the University of Westminster and others. On the other hand, Microsoft claims that Live@edu is the most widely used cloud productivity service for education. With the switchover to Office365, the offer includes everything available in Office 365 for enterprises, helping teachers save time and manage their curricula while giving students access to tools that make learning more inspiring, relevant and collaborative.

Besides this typical SaaS adoptions by the education, the next popular cloud implementation is starting a private IaaS cloud in order to increase the effectiveness of the available lab equipment and turn the physical computers into virtual laboratories. One of the most successful stories is the Virtual Computing Laboratory (VCL) [4], [5] developed with IBM for the uses of the University of North Carolina. The main objective of the project was to move the resource from the level of the individual towards the level of a complete laboratory or a synchronous or asynchronous learning classroom. The resources needed for accomplishment of the lab or class tasks should be available on demand for the time period of few hours to few weeks. The basis of this solution is an IaaS service model with storage of a number of different VM snapshots needed for different tasks. On the top of the virtualized infrastructure there is a specialized broker that responds to the users request and combines different snapshots when needed.

The cloud computing potential for efficiency improvement, reducing costs and a number of other benefits is recognized by a number of educational institution in USA, Europe and beyond [6], [7]. The successful cloud computing in the education stories also include the underdeveloped regions of Africa where the use of Google Apps and Microsoft Azure is rising rapidly. By the means of donations from the World Bank and other financial institutions, the teachers and students in the educational institutions have the opportunity to obtain the minimal infrastructure needed in order to obtain access to the cloud services.

All investigated examples have highlighted the benefits of using cloud services, but have also pointed out some challenges and concerns that include risk assessment, security, and governance issues, as well as uncertainty about return on investment and service provider certification; and questions regarding which activities are best suited for the cloud.

Thus, while the educational leaders are acclimating to the new IT environment and are testing the marketplace with ventures ranging from computing cycles and data storage to student e-mail, disaster recovery, or virtual computing labs, many remain cautious observers as they assess the potential impact. This points out to the necessity of careful and in-depth analysis of the decision making towards the migration to the cloud [8][9].

III. Towards Migration of Education in the Cloud

Education institutions must determine what makes sense to move to the cloud and why. This is important from a risk standpoint as well as from a cost perspective. For instance, many have already inclined to free e-mail commercial providers, using the commercial provided cloud for knowledge assessment may prove a bad idea if a service disruption occurs during a peak testing period. Thus, security levels around all relevant data should be considered and service levels must be guaranteed. Of course, one important issue that must not be forgotten is to identify the options for exiting the relationship with a commercial provider or switching to a different one.

In general, it is advisable to make the migration process a step by step venture, where the start is paved with the obvious choices that bear low risks like web applications with public content that can benefit from the scaling in the cloud.

Based on past experiences [10], we can sublime the necessities for cloud services in the education as given in Table I. Basically the services offered here are making up a could supported learning management system that is course oriented.

TABLE I. Educational Services in the Cloud

Service type	Service description
Cloud Courseware	Enables the students to acces the coursework in digital form
Cloud Classroom	Interactive teaching independent of the teacher or student location
Cloud Lab	Remote access to virtual computers, modern equipment and software, esspecially important for technical courses
On Demand Lessons	Enables students to demand desired courses
Student assignments	Enables student to download homework assignements and upload the solutions
Online evaluation and feedback	The students can do an online knowledge evaluation and get corresponding digital certificates
E-Publishing	Servise that enables the teaching staff, students and other users to access useful electronic publications and create web sites
Data sharing	Sharing information when working on team projects, research assignements, etc.
Remedial classes	To help students that can not attain the required knowledge level just by following regular classes

Other systems that can extremely benefit from moving into the cloud include a resource booking system, support/ticketing system, alumni/membership manager and dashboard with a single sign-on.

In order for the services to be properly used and secured a robust identity and access management is a must. On this point, a key technology that is to be used are the federated identities, the ability of users authenticated in one domain to access resources elsewhere. Furthermore, with the special regulatory concerns about student data, educational institutions have particular need for an identity ecosystem that must be recognized by the commercial cloud provider. It is also expected that policies must be developed for handling the increasing popular user-controlled web-based identifiers, like those offered by Google and Facebook.

On the other hand, all of these concerns are shedding light on the nature of the cloud in which the majority of the education related services could be migrated. Due to the extremely sensitive nature of some of the data, the cost effective option is the hybrid cloud. In the hybrid cloud environment, one or more commercial cloud provider can offer the low sensitive applications (like access to coursework, student data shares) from within the public cloud, while the high sensitive data and applications can be hosted inside a private cloud that will offer the needed security level while leveraging scalability.

We recommend the seven-step model [11], presented in Table II, for migration to the cloud that should be implemented in several phases in order to ensure smooth transit for the existing applications that can benefit from the cloud.

TABLE II. SEVEN STEP MODEL FOR MIGRATION TO THE CLOUD

Step	Description
Assessment	Assessment of the issues relating to migration, at the application, code, design, and architecture levels
Isolation	Isolation of all the environmental and systemic dependencies of the enterprise application
Mapping	Separate the components that should reside in the captive data center from the ones that will go into the cloud
Re-architect	A substantial part of the application has to be re-architected and implemented in the cloud
Augment	Augment the apps with the features of cloud computing services
Testing	Testing and validation of the new cloud apps
Optimization	Iteration and optimization based on the test results

The phases will be made up out of groups of new and/or existing applications classified on the basis of complexity, security issues and data sensitivity, so that the lessons learned from the non critical groups can be implemented thus increasing the quality and ensuring success with the risk prone services.

IV. CLOUD EDUCATIONAL SERVICES CONCEPT FOR MACEDONIA

Since the educational process, especially in the elementary and secondary education, is an integral part of the wider community, it becomes one of the mostly influenced on fields that have to keep locked step with the current technological trends. On the other side, on of the main benefits of cloud computing is its cost effectiveness. Thus, it is imperative that the educational system in Macedonia governed by the Ministry of Education and Science (MON) envisions its current and future projects implementations in the cloud computing environment. As a first step towards this process, in this section we give an overview of the current status of the educational facilities in Macedonia and discuss the possible design of a cloud integrated solution.

A. Current Status

The state schools in Macedonia have a very similar IT infrastructure and can thus be easily encompassed with one integral solution. Within the “Computer for each child” project [12] since 2006, each school is equipped with modest performing PCs for the students that are interconnected in a local network with an Internet access. For the needs of the students a set of educational apps are pre-installed on a Linux based OS. As part of the same project, each teacher is given a small laptop with low performances. During the “modernization of education” project [13] the schools were able to update some of their equipment, but this budget was mainly being spent on LCD projectors. The schools administrative and management staff on the other hand are usually equipped with a more up-to-date infrastructure and usually have a separate connection to the Internet. All of these IT equipment is being managed by a small number of part time regionally assigned administrators that are tasked with hardware maintenance and occasional software updates. However, praxis has proven that daily support and management is needed. This hardware review of the current status reveals that a cloud solution could be a very promising solution to the problem of constant update and management of the equipment.

As for the current software status, there are several currently active projects developed and supported by the MON. On the state level an Education Management Information System (EMIS) [14] has been implemented since 2009 with an included module for Human Resource Management. The system is web based with an aim of keeping current data on the school employees, with their qualifications and work positions, as well as coarse grained data on the number of students, classes, work plans and other teaching activities in the school. Another active project is the e-Journal [15], that aims to enhance the communication between the teachers and the student parents, by allowing a web based access to the parents into the daily journal of teaching activities and the student progress expressed via the trimester grades he/she receives for each attended class. It also provides statistical data for the Ministry and other relevant institutions. This web-based application is connected to the EMIS system in order to ensure data integrity for the overlapping data.

The most current addition to the educational projects is the web based integrated e-testing information system [17] under the control of the state-testing center. The aim of the system is to provide a standardized external testing environment that will ensure an unbiased external testing with a relevant global state scale results. The system has been functioning since 2013 and is supposed to be used only in parallel all over the country in strictly previously defined testing periods.

A recent, cloud related, project is the Live@edu initiative that has been piloted in 2010 in a number of preselected schools. As a part of this project the schools are able to design the schools web site based on Microsoft SharePoint with a free one year hosting and training of one web administrator. The major benefit of the project is the free use of the Microsoft Office 365 suite in the cloud, supported with only 20 emails with 10GB inbox, 25GB cloud storage on OneDrive and possible video chat with other pilot schools via Lync. This project represents a small first step towards the cloud initiative and in order to show real benefit for the schools its offers need to be expanded so that all teachers and students can create a digital identity (liveid in this case) with possibilities for sharing digital data via online storage and emails and communicate based on a global/local directory.

Another parallel initiative is the implementation of the open source e-learning platform Moodle by a few of the schools in the country, which has proven to be a bit of a problem for wider acceptance because of the need for hosting and management of the services in-house on the schools scarce resources. A last active project example is the media library that is created with the intent to provide the students with a multimedia material (mostly educational videos) that will enhance their learning process. The videos are actually materials already present on the Internet, preselected by the relevant teaching staff and are made publicly available.

As previously stated, the available IT hardware at the schools is difficult to manage and even more difficult to update over a prolonged time period. In addition to that, the software applications already in place are very good candidates for migration into the cloud in order to improve their reliability and performances especially in critical moments (like graduation testing or trimester grading) when it has been shown that the systems are experiencing difficulties due to the number of parallel active users. The live@edu project is a good starting point for cloud based office apps, calendar, email and storage access as well as collaboration based on the SharedPoint features. However the implementation is confined on the school level while a global approach can offer far more benefits.

B. Migrating the Services in the Cloud

The first step towards one cloud integrated solution would be to identify all of the participants in the educational process and devise the needs, activities, access and security levels for each of them. In light of this we must emphasize that the current services do not include the main participant: the student himself, which is a major setback that disables the schools to keep up with the evolving teaching methodologies. Thus, an effort for an integrated, yet modular, cloud oriented solution that encompasses the students, teachers, parents, administrative and management staff, relevant associations, as well as the relevant state bodies should be included. Also, the usability and friendly design should be targeted for the most numerous users of the system: the teachers and their students.

The idea of a hybrid cloud with a private cloud managed by MON that can host the sensitive SaaS services like EMIS, E-journal, e-Testing and alike, and the public cloud supported by a commercial cloud provider can offer hosting of the less sensitive SaaS based services like the courseware, classroom, remedial classes, labs, shared data, integrated communications and alike.

A major part of the public cloud should be given to a VCL like solution, or in the least an IaaS services provider with a set of predefined snapshots. This setup will enable the students to have an access to a standardized virtual desktop. Using this model MON can retain the control over the chosen applications and services that will be used in the daily school activities, while the planning, securing, management and maintenance of the hardware infrastructure are given to the provider. The provider must ensure scalability and QoS, while at the same time the great number of students will ensure a low cost solution based on VMs especially when compared to the constant need for hardware upgrade and maintenance which will needed in a very less frequent intervals since the existing equipment is more than adequate to act the part of a terminal for the students virtual desktops.

The rest of benefits of the virtual desktop is the possibility to deliver it to a wide range of different devices like desktops, laptops, tablets, smartphones. Each student can own his personal preset VM that can suspend and activate upon request. This will not only allow the students to access a standardized environment from their homes as well as from the school, but will also work towards the future trends of the schools in the sense of the “bring your own device” possibilities. Another possibility for implementation is to enable an enhanced virtual desktop supported by the provider (this elevates the service to a Desktop as a Service - DaaS level [18]), where the provider can deliver only a requested application from the snapshot along with a virtual storage for the active data.

Also, one must not forget the cost effectiveness of the possibility to downscale the public IaaS cloud resources during the winter and summer breaks. Of course, the most imperative in this scenario is the control over the VM snapshots that has to be implemented by MON in cooperation with the relevant institutions. A dedicated team that will manage the pool of snapshots and monitor the resource usage needs to be assigned. We must stress that the teachers will also use their own snapshots with additional preset applications that will enable them to manage the rest of the classroom and collaborate with the students. Special care must be given to the chosen applications and SaaS applications design in the light of user interface consistency so that the users have a seamless and effortless experience while using a blend of the VM environment and the browser based SaaS applications.

The benefits that can be obtained with the setup of the private cloud inside MON and migrating the existing active web based to this cloud hosting are apparent. All of the previously mentioned applications: EMIS, e-Journal, e-Testing; are perfect candidates for the cloud migration. This is especially the case with the e-Testing application since it has a very irregular pattern of usage. Very low usage some time before the overall state testing of students for a given course material. In the current state, the state has had to provide the hardware infrastructure that will ensure no problems with the application performances during the testing, and this hardware remains severely underutilized in the rest of the period.

Hence, if one gathers all hardware used to host all of the afore mentioned services, it can be used as a setup for a private cloud datacenter in which these services will be hosted. This setup will enable a lot higher utilization of the resources and increased performances of the applications. Having in mind that the database for most of them is already shared to somewhat extend, this reconfiguration is a natural step in the process in order to increase the efficiency and effectiveness of the available hardware. With this the private datacenter will also be able to host additional services that can be developed in the future without or with a minimum need of hardware investment. Another option that can be considered is extending the processing power of the private cloud with a commercial cloud support in the moments of high demand. It is imperative that a very well devised backup and disaster recovery solution exists for this private data center, where one of the options is a regular backup to a specialized backup cloud provider. In this case, the contract with the cloud backup provider must clearly state where and how the sensitive data will be stored in order to satisfy all legal and security issues that may occur.

One important part of the cloud solution must be the cloud storage option, which can be set up in the public part for all non-sensitive data that are restricted by any legislation. These can include online storage for public data of the schools, for work class materials, educational multimedia and alike, as well as shared storage for class assignments and similar activities. The sensitive data that contains student records, e-testing results and exams and similar should be kept in the private cloud. However, we would like to bring to attention the importance of choosing the right public cloud storage from several different aspects like compatibility and integration with the existing applications and services, but most importantly real privacy and protection of the data at the provider side. For an example, one of the most popular cloud storage providers today Dropbox keeps the users files on his storage servers unencrypted and demands right over them in the given terms of use unlike his competitor Box [19].

C. Potential risks

The proposed educational cloud solution has one potential weak point in the current setup. This is the increased necessity for Internet traffic from and to every school. The current Internet connection bandwidth provided to the schools is with around 9 Mbps for download and 0.7 Mbps for upload. However, this creates problems for even the todays normal functioning within the school activities and needs to be upgraded soon independently of the cloud initiative.

With the aim to setup a cloud apps friendly Internet connection for the schools and other relevant institutions, while doing the upgrade one must have in mind that one of the changes in the Internet traffic that the rise of the cloud services has induced is the symmetrical traffic [19]. Thus, during the upgrade it is imperative that the future Internet connections of the schools have not only wider bandwidth but also symmetrical upload and download.

D. Criterias for Choosing the Right Provider

According to the proposed solution a list of criteria needs to be compiled in order for it to be presented to the potential providers. The most important criteria that have to be included are given in Table III.

TABLE III. CLOUD PROVIDER CRITERIAS

Criteria	Description
Functionality	A list of all needed functionalities like apps, storage space, types of collaboration and communication, etc. Special attention must be given to requirements for cross compatibility especially if more than one provider is chosen
Platform	In order to ensure a consistent user friendly interface and work environment
Technical issues	Workflow and timeline for migration of the existing applications into the cloud Option for automatic profile creation and creating federated identities System management and administration
User access and user experience	The level of complexity of the services provided must be adapted to the students age, but options for talented students must be also provided Solutions for people with special needs
Price	Clear definition of the payment process together with a bill management automatization according to agreed rules of payment. Special attention must be given to the types of resources involved and pricing of the traffic to/from the cloud provider
Test period	The fulfillment of the contract must be based on the delivered quality of the service, which for the applications has to be evaluated during a testing period. The workconditions during the testing must be equal to the typical usage of the service.
Optimization	Iteration and optimization based on the test results

Drafting the agreement with the cloud provider may easily become one of the most crucial steps in the whole process. Basically there are two types of agreements possible: predefined and negotiable. The predefined contracts are the basis of the scaling economy and are one of the major benefits of the public cloud. In this case the provider is the one that gives all terms and conditions of use, and usually the client is not protected in case of changes. This is of course unacceptable in this case, and thus a negotiated contract is needed. A negotiated contract will include the needs and terms of the client side especially regarding the data security, privacy, encryption, application isolation from other cloud users, service monitoring, technical control, exit strategy, legislation conditions and other. Because this part is crucial in the whole process it is advisable that a third mediation party, i.e. broker is enlisted. The broker will ensure that all client demands are fulfilled including detailed service description including actual location and security measures; policies, standards and procedures that are going to be applied; type of connection to the provider; availability, reliability and scalability; service level agreements; disaster and recovery options; and others.

V. CONCLUSION

Cloud computing has brought forth revolutionary changes in the IT industry. Instead of buying expensive hardware, software with further maintenance expenses, the Internet users can now buy what they need, when they needed and pay accordingly. The cloud service providers are offering a wide variety of platforms and services delivered to the big and small customers from their specialized cloud data centers.

Educational institutions from all over the world are becoming a part of this cloud revolution. Examples range from outsourcing e-mails to a cloud provider, to taking the whole educational process on a whole new level by exploiting the cloud benefits.

In this paper we have given a proposal for a possible integrated cloud solution of the elementary and secondary educational services in Macedonia. The proposal has been made based on the current status and the current active educational project, while also drawing from the existing experience presented in the reviewed example.

Our final goal was a complete cloud based system that revolves around the most important participant in this process, the student.

ACKNOWLEDGMENT

This work has been partially supported by the EUROWEB project (http://www.mrtc.mdh.se/euroweb) funded by the Erasmus Mundus Action II programme of the European Commission.

REFERENCES

[1] K. Tejaswi, "Cost effectiveness in Educational institutions using Cloud Computing," (2013.

[2] Microsoft cloud computing in education, Live@edu project, http://www.mtek.mk/2011-12-12-11-08-01/proekt.pdf

[3] Google Apps for Education, https://www.google.com/enterprise/apps/education/

[4] M. A. Vouk, “Cloud Computing – Issues, Research and Implementations”, Journal of Computing and Information Technology - CIT vol. 16, 2008, no. 4, pp. 235–246 doi:10.2498/cit.1001391

[5] Virtual Computing Laboratory, http://vcl.ncsu.edu

[6] N. Sultan, “Cloud computing for education: A new dawn?,” International Journal of Information Management vol. 30, 2010, pp.109–116

[7] Cloud computing in education, UNESCO Institute for Information Technologies in Education

[8] J. Powell, “Cloud computing – what is it and what does it mean for education?”, Leicester Business School, DMU

[9] M. Patra, M. Ranjan, and R. K. Das, "CeMSE: A Cloud enabled Model for Smart Education," 2013.

[10] Wholeschool, the go-to for education, http://www.wholeschool.ie/Home/About

[11] R. Buyya, J. Broberg, A. Goscinski, “Cloud computing : principles and paradigms”, John Wiley, 2011

[12] “Computer for every child” project of the R.M. goverment http://vlada.mk/proekti/kompjuter-za-sekoe-dete

[13] E-government strategies for R.M, http://www.mio.gov.mk/files/pdf/dokumenti/Strategija_za_e-Vlada.pdf

[14] “Modern education” MON project http://www.mon.gov.mk/index.php/aktivnosti/627-2010-07-23-09-51-59

[15] Education management information system -EMIS http://www.emis.mon.gov.mk/

[16] E-journal http://ednevnik.edu.mk

[17] Integrated e-testing information system, http://www.eet.mon.gov.mk/images/userManuals/UserManual.pdf

[18] J. Liu, H. Yan, C. Zou, H. Suo, “Architecture of Desktop as a Service Supported by Cloud Computing” Advanced Technologies, Embedded and Multimedia for Human-centric Computing Lecture Notes in Electrical Engineering Vol. 260, pp 355-361, 2014.

[19] Y. Zhang, C. Dragga, A. Arpaci-Dusseau, R. Arpaci-Dusseau, “Box: Towards Reliability and Consistency in Dropbox-like File Synchronization Services”, 5th USENIX Workshop on Hot Topics in Storage and File Systems, 2013.

[20] R. Bianchi, G. Herrmann, “Cloud Services need Symmetrical Data Rates: enhancing the Network-Compatibility of SHDSL.bis and ADSL2+”, Kommunikationskabelnetze - 18. ITG-Fachtagung, 2011

Session 5

eWorld – eWork, eCommerce, eBusiness, eLearning 1

An Overview of the Systems of Keyword Search in Relational Databases

Ivan Bislimovski
Student at Faculty of Computer Science and Engineering,
Computer Science department
Skopje, Republic of Macedonia
Email: ivan.bislimovski@arhiv.gov.mk

Goran Velinov
and Margita Kon-Popovska
Faculty of Computer Science and Engineering,
Computer Science department
Skopje, Republic of Macedonia
Email: goran.velinov@finki.ukim.mk
and margita.kon-popovska@finki.ukim.mk

Abstract—**Keyword search in relational databases is useful and helps users to search without any database technical background. Keyword search is widely used on the Web, on the other hand, relational database is dealing with structured data and assists users to query information using structural query language (SQL). With the increase of the amount of data stored in relational databases the need for using keyword search is increasing. However, efficiency and effectiveness (finding right information) still remain a problem. In this paper, several systems that use relational database with keyword search are examined. We overview, analyse and compare the functionality of the systems, emphasizing the strengths and weaknesses of each. Finally, we discuss the most important recommendations for the keyword search process, and propose the best features such a system should have.**

Keywords—Keyword Search, Relational Databases, Analyse, Evaluation.

I. Introduction

In the last several years the interest for the keyword searching is increasing. The web browsers are internet pyramid based, searching large amount of information and have users from all around the world. The regular user has access to almost all information using internet connection.

Internet search engines introduce the keyword search. On the other hand traditional Relational Database Management Systems (RDBMSs) provide powerful query language that does not involve keyword search. Certainly, large number of tuples stored in database systems can be searched using complex SQL query but could not be easily understood by all users. With the introduction of the keyword search in relational databases, the need users to know the database schema will be avoided. Keyword search is working by keywords matching the values in tuples of the attribute that is associated with multiple relational tables.

Today the relational databases are widespread and they are the main technology for information storing. Most of the web services or applications and systems are accessing the content using database management system. With the increasing of complexity and amount of data in databases, the need of keyword search based retrieval is growing.

Example

Text, semi structured and structured data are often stored using relational database management system. In following example, archive collection database table, Clt(fondID, boxNum, arhINum, paperFrom, paperTo, description) enables users to write comment about a particular collection (attribute, description).

```
SELECT *
FROM Clt AS C
WHERE CONTAINS(description, ’completed’, 1) > 0
ORDER BY fondID(1) DESC
```

Keyword query returns tuples from the table Clt that match the word ’completed’, listed in descending order by fondID. The tuple values from the collection measure the column description that suits the above query. Often it is necessary to provide a good solutions from the keyword query concerning tuples from multiple tables. This means that there is need to join multiple tables and implement additional semantics.

Database management systems allow use of queries that contain text values but also recognize the importance of the ranking strategies. With search functionality shown above, we specify the tuples that exactly match the keywords from a given query. Often we need ranking list of tuples that by some criteria best match the keyword query, usually not knowing table structure.

Due to the increased use of keyword search approach, there is a need of development of keyword search systems using relational database. There are several existing systems which can be used in this field. Most of them are academic projects such as: Spark [1], Discover [3], DbXplorer [4], Banks [5], Blinks [6]; Commercial projects such as: Spark2 [2] can also be encountered.

Generally, our challenge is to examine all the details of the keyword search process, techniques used in developed systems, and their basic components. So, in our knowledge no prior paper exist that provide a complete overview of keyword search process. Our aim is the overall evaluation of the quality of systems by comparing their functionality and differences, by analysing the structure of the systems, by statistic evaluations

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University.

of their performances, to ensure enhancements and future progress in their efficiency and effectiveness.

Organization

Section 2, gives overview of existing systems. In section 3 we present a comparison of the systems using systems characteristics. Section 4 outlines the conclusions and recommendations.

II. Overview of the Existing Systems

Keyword search process can provide many new opportunities to the relational database and also simplify the user requirements. In this section we overview the keyword search prototypes in relational databases by description of common system structure of several keyword search systems.

A. System structure

The systems can be briefly divided into two categories: relation based systems and graph based systems. In relation based approach the answer to the keyword query is a set of tuples that contains all or some of the search keywords. With each primary-foreign key relationship, the selected tuples can be joined together. These systems can be used for all data types and the result of a keyword search usually connects the tuples corresponding to the keywords. On the other hand systems that belong to the graph based approach consider relational database as a graph, which contains set of nodes and set of edges. For two given nodes there is an edge in the graph, if primary-foreign key join relation exists. Graph based approach is used where entire content in the relational database is presented as a graph, where each tuple is presented as a node and each primary-foreign key relationship is presented as an edge. In these systems keywords given by the users are used to generate the tuples using an approach similar to breath first search (BFS).

B. Existing Systems

Predominantly all the systems use relational database and web server architecture that allows users to interact with system using web browser. Mainly, all systems use architecture using relational databases. With this architecture, it is clear that users do not need to install any specific software.

The search technology and utilized ideas have their origin in academic research environment and has been the subject of many industry publications. To build and develop such systems, it is necessary to optimize, rank, index and choose algorithms that will bring optimal and desirable results.

SPARK

The SPARK [1] system allows limited query capacity and flexibility using a few keywords for searching, usually only two keywords. The system allows joins that correspond to the primary-foreign key and are matching the keywords from the given query. Afterwards, the system uses an effective method for ranking the answers. The system uses a simple interface. Firstly, the user chooses one of the search modes (standard or advanced). Secondly, the user selects a data source and inputs a keyword query. The advanced method additionally allows to search by phrase ( ), by wild card (*, .) and by schema from the database.

SPARK2

The SPARK2 [2] system, works very similar to the ”Google” search engine using full text search across relational and structural data. Unlike the traditional web search, the SPARK2 system uses search methodologies that respond to collections of predefined data. The search focuses on predefined relationships and dynamically constructs data by connecting mutually referring pieces of information that are stored separately. The system creates a process of setting up the relational database and almost always provides the desired result. The SPARK2 proposed the skyline sweeping algorithm that achieves minimal database probing by using a monotonic score upper bounding function for ranking formula. The key idea is to check only the necessary candidate pairs where upper bound score function is higher than the unchecked candidates. Another, more effective algorithm is block pipeline algorithm decomposing the entire search space into blocks, such that there exist a monotonic upper bounding function that bounds the score for all candidates. The block pipeline algorithm can achieve greater orders of speed-up comparing previous algorithms.

DISCOVER

The DISCOVER [3] system use the schema of the relational database management systems. The approach leads to effective algorithms that answer the keyword query, because the structural information contained in the schema perform the query processing. DISCOVER system is based on a simple architecture and use Boolean-AND semantic query. The DISCOVER system (as also the BANKS system) require all keywords search from the query to emerge in the tree, as nodes and tuples that can be returned as a result of the given query. Then we are creating ranking techniques based on the size of the result.

DBXplorer

The DBXplorer [4] system is similar to DISCOVER [2] system architecture. It also uses Boolean-AND semantics to return the tuples from single or multiple tables via joins. DBXplorer is using Microsoft SQL Server 2000 as database server, Microsoft IIS as web server and ODBC for communications with the database. DBXplorer builds a join tree for keyword found in multiple tables. For each join tree, SQL statement is build. Results are ranked based of the number of joins and the tuples that have fewer joins are ranked higher.

BANKS

The BANKS [5] system comprise the problem of keyword search using structured and semi structured data. The BANKS system has a graph view on the database, where the tables from the database are represented as nodes, while relationships between the tables are represented as edges. Edges describe the join relationship established between the primary-foreign key relations. The system corresponds to a query containing all the searching keywords using Steiner trees [5] and is using heuristics during the search process. Goldman et al algorithm [5] uses related views based on the graph structure to view the database. The user query specifies two sets of objects, set of discovered objects and set of near objects. The system then uses a set of ranking objects found by comparing their location. If the result is not complete by discovered objects, near objects are also included. Simple search algorithm is used. The algorithm calculates the efficiency of location with placing hub indexes. Finally, after we generate the graph the important structural tuples taken from the database are displayed to the user.

BLINK

The BLINKS [6] system is based on a data graph structure which is used in multiple systems. Top-k keyword query for search graph finds the top-k results by using a specific strategy for ranking where each answer is structured under the graph containing all the keywords. To get better results, the BLINKS system uses bi-level index for searching keywords from the graph and it gets simpler and quicker search result. The system divides the data into blocks, bi-level index stores the most important information in the block in order to establish a search through blocks.

III. Comparison and Discussion

A. Comparison

Following we are comparing the most relevant features of the above described keyword searching systems.

The DISCOVER and DBXplorer keyword search systems generate tuple trees from the multiple relations in the database to identify the tuple trees that match all the keywords from a given query. The tuple trees are minimal number of candidates, if the more than two keywords are used, tuple trees are used. The BANKS system allow more associative criteria such as allowing keywords to be part of the metadata of the database. The SPARK and SPARK2 can efficiently locate matching tuples for each search keyword and form the non-free tuple sets. In BLINKS system, the database is presented as a graph and tuples are treated as nodes connected using primary-foreign relationship. The BLINKS uses backward search strategy which explores the graph starting from nodes containing query keywords. Also the systems produce candidate networks (CN) [7], [8], such that the corresponding query may produce some query result. The DISCOVER and DBXplorer systems have similar structure and use ranking strategies, where ranking is based on the number of joins involved in the tuple trees. More joins meaning greater possibility of comprising all the keywords for a given query. On the other hand, if there is high number of joins, the result might not be helpful as we seek. BANKS system uses weights which is similar to PageRank for websites and is weighting each edge in the tuple tree. The BLINKS system uses sorted function that focuses on indexing and query processing. The sorting function from BLINKS incorporates several features from IR-style technique [11], [12].

The systems DISCOVER, DBXplorer and BANKS systems do not use any IR ranking characteristics. According to several researches [10], [11], [12], the IR technique have been significantly successful for ranking. The outmost created systems such as SPARK, SPARK2 and BLINKS use the IR-style technique for more efficient ranking strategies over text. The IR community is focused in improving the efficiency issue and quality of ranking functions. On the other hand systems that use different strategies, ignore some important factors that are critical for keyword search process.

When the value of the attributes is text and the text contains multiple keywords, obtaining the columns that contain the keywords can be done by creating the substring that matches the text. In this case systems use full text search that is more effective and faster comparing with the LIKE statement that is slow in a case the search starts with wildcard. Some systems can allow creating AND-OR semantics for the keyword query.

Architecture

Mainly, all systems use architecture using web server and web browser. With this architecture, it is clear that users do not need to install any special software. Users need only a web browser that accesses the service and mostly because of these features we use internet search. Therefore, for the easier functioning, the system needs to use the appropriate database and web server that communicates with the database using the specified standard. Systems can additionally allow to be used internally via intranet and it can perform search from selected databases.

Performance

If the top-k searching does not upgrade performances of the keyword search, it is necessary for the systems to use optimal ranking algorithm. On the other hand, while the system produces more solution answers to the query, for some of them it may take a very long time or it cannot produce a solution answer due to lack of memory capacity. These performance problems must be avoided when we are working with real world databases.

Sometimes information would be too large to store and too expensive to maintain for large databases. Some of the systems used different type of indexing which accelerates the computation and increase search speed.

Ranking

If the expected number of answers in keyword search is too big, various mechanisms - algorithms for ranking the answers are introduced.

Existing ranking algorithms can generally be classified into four categories [13]:

- Link-size based ranking algorithm (DBXplorer [4] DISCOVER [3])
- Weight based ranking algorithm (BANKS [5])
- Information retrieval ranking algorithm (IR Style [11], [12], SEEKER)
- Authority-transfer based ranking algorithm (ObjectRank)

Query Processing

In the keyword search the last and most critical process is query processing. The efficiency of query processing is vital part when we are dealing with large database. The query processing time is large for large databases, if query processing is not fully optimized. Each system use algorithm that implements the query differently, but it is important that each algorithm performs a complete optimization of ranking function and top-k query.

In the following section focus is on the comparison of the systems features and their important components. These elements raise additional challenges and they remain an open research topic in order to improve entire keyword search process.

C. Recommendation

To improve the keyword search process, we recommend considering the following components.

Algorithms

Usage of the appropriate score function can modify the existing algorithms [12] so that it will calculate the desired answers more efficiently. Since the algorithms can contain many unnecessary join checking, the recent algorithms [2], ensure that these checking do not occur and that accesses to the database are minimized. The latest systems have better algorithms and ranking strategies and are designed to reduce the cost that typically dominates in the executing. Improvements and creation of better algorithms, is challenge that can further reduce the excessive amount of work.

Ranking function

Ranking functions using IR (information retrieval) techniques [10], [11], [12] are used. According [2], [12], ranking functions used in above systems are non-monotonic in order to obtain more efficient score. It is still an open question how a ranking function can obtain better query results.

Query processing optimization

The query processing time is large for large databases, if query processing algorithms is not fully optimized for ranking function and top-k queries. The challenges of query processing mainly lies in using non-monotonic score function. This function is used in the skyline sweeping algorithm [2] and block pipeline algorithm [2]. Previously none of the existing algorithm has used top-k query processing methods to deal with non-monotonic score function. For each candidate network, each probe involves complex query and joins from different relations. This thread is more studied and observed in most query processing engines.

Keyword Search versus Structure Query Language:

One of the aspects for effective query is to use calculating factor for top-k results. Top-k query has a problem with finding the best answer for the specified query. The answer is determined by the ranking function to set a score individually for each query. If there are k answers for the given query, the user has a maximum choice, but not complete result. Keyword search is useful when user is unfamiliar with the database schema and then user can use keyword search technique to obtain a result that contains most of the desired data records. On the other hand, structural query language obtains more strict and precise result. The user knows which tables and attributes to query to obtain the result. Therefore, the keyword search in relational database should not be used as a replacement tool for structure query language, it should be consider as an additional tool that helps the user to get wanted result with less knowledge of database structure.

IV. Conclusion

The amount of information stored in the relation databases increases, therefore the need for efficient retrieval of the information also increases. Keyword search allows users to search information without having knowledge of the database schema and SQL query language. In this work several existing keyword search systems (SPARK, SPARK2, DISCOVER, DBXplorer, BANKS, BLINK) based on relational databases are examined. Their most important components, functionalities and features are overviewed, analysed and compared, emphasizing the strengths and weaknesses of each.

As the analyze result best features of the existing systems were highlighted, stressing the features that each keyword search system should have. Also several important recommendations for the keyword search process and future improvement of the systems were recommended.

Work can be considered as starting guide to researchers to better understand the entire keyword searching process using relational databases. A critical analysis of this process is useful because there are significant differences in efficiency and effectiveness between existing systems. Knowing strengthens and weaknesses can eider help in choosing the appropriate system or in developing better one.

References

[1] Y. Luo, W. Wang, and X. Lin. Spark: A keyword search engine on relational databases. *ICDE*, 2008.

[2] Y. Luo, W. Wang, X. Lin, X. Zhou, J. Wang, and K. Li. Spark2: Top-k keyword query in relational databases. *TKDE*, 2011.

[3] V. Hristidis and Y. Papakonstantinou. Discover: Keyword search in relational databases. *VLDB*, 2002.

[4] S. Agrawal, S Chaudhuri, and G. Das. Dbxplorer: A system for keyword-based search over relational databases. *ICDE*, 2002.

[5] B. Aditya, G. Bhalotia, S. Chakrabarti, A. Hulgeri, C. Nakhe, and P. S. Sudarshan. Banks: Browsing and keyword searching in relational databases. *VLDB*, 2002.

[6] H. He, H. Wang, J. Yang, and P. S. Yu. Blinks: Ranked keyword searches on graphs. *SIGMOD*, 2007.

[7] J. Xu Yu, M. T. Ozsu, L. Chang, and L. Qin. Keyword search in databases. pages 1–81, 2010.

[8] J. Xu Yu, L. Qin, and L. Chang. Keyword search in relational databases: A survey. 2010.

[9] Daniel P. Sokol. Keyword search in relational databases. 2004.

[10] W. Wang, X. Lin, and Y. Luo. Keyword search on relational databases. *NPC*, 2007.

[11] F. Liu, C. Yu, W. Meng, and A. Chowdhury. Effective keyword search in relational databases. *SIGMOD*, 2006.

[12] V. Hristidis, L. Gravano, and Y. Papakonstantinou. Efficient ir-style keyword search over relational databases. *VLDB*, 2003.

[13] C. Wang, J. Ding, and B. Hu. Ranking algorithms for keyword search over relational databases. page 1, 2012.

Child Online Protection Campaign – Macedonian Experience

Smilka Janeska-Sarkanjac
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
smilka.janeska.sarkanjac@finki.ukim.mk

Ivan Chorbev
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
ivan.chorbev@finki.ukim.mk

Vesna Dimitrova
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
vesna.dimitrova@finki.ukim.mk

Gorgi Madzarov
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
gjorgji.madzarov@finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
dimitar.trajanov@finki.ukim.mk

Abstract—Internet, as a disruptive technology, changed numerous things in the society. Entire industries disappeared. New business models were invented. It affected the way we communicate – the information we exchange is more like tweet or texting than like 19th century novel. Internet opened unprecedented communication channels: Facebook, Twitter, WhatsApp, Instagram, Pinterest, Skype, Youtube. People are networked and interconnected.

The other side of the story is that Internet facilitated crimes also: impersonation fraud became identity theft, copyright violation became file sharing, and accessing censored materials—political, sexual, cultural—became trivially easy.

One of the most vulnerable socio-demographic cohorts on Internet are the children. International Telecommunication Union started global initiative named Children Online Protection, addressing legal, technical, organizational and procedural issues as well as capacity building and international cooperation. Macedonia took part in this effort. The Ministry of Information Society and Administration of the Republic of Macedonia, the Faculty of Computer Science and Engineering, and several other organizations and companies organized number of events, lectures, workshops and provided free browser for the children that controls Internet access. One of the outcomes of these activities was survey conducted among parents and guardians regarding behavior on the Internet of the children and their guardians, and their perception of the Internet security and undertaken protective measures.

This paper describes the activities conducted within this initiative in Macedonia, and analyzes the results of the survey.

Keywords— Online security, children vulnerability, children online protection, Republic of Macedonia

I. Introduction

The Child Online Protection (COP) Campaign was conducted in the Republic of Macedonia during the 2013. The initiative originated from the Ministry of Information society and administration of the Republic of Macedonia, and was supported by a number of public sector institutions, business companies, and media sponsors: the Government of the Republic of Macedonia, Ministry of education and science, Ministry of labor and social policy, Agency for electronic communications, mobile operator company One, daily newspapers Dnevnik, Vest, Utrinski vesnik and Sport, and the Faculty of Computer Science and Engineering, Ss Cyril and Methodius University in Skopje.

The campaign was part of global Child online protection initiative, organized by ITU (International Telecommunication Union). "COP aims to tackle cybersecurity holistically, addressing legal, technical, organizational and procedural issues as well as capacity building and international cooperation" - as it is stated by ITU [1].

The global Child online protection initiative had four fields of action, and the same number of guidelines were produced: guidelines for children, for their Parents, Guardians and Educators, for the industry and for the policy makers.

The purpose of the Macedonian campaign was to raise the awareness for the online threats that children face, and to improve the level of security of the children when using the

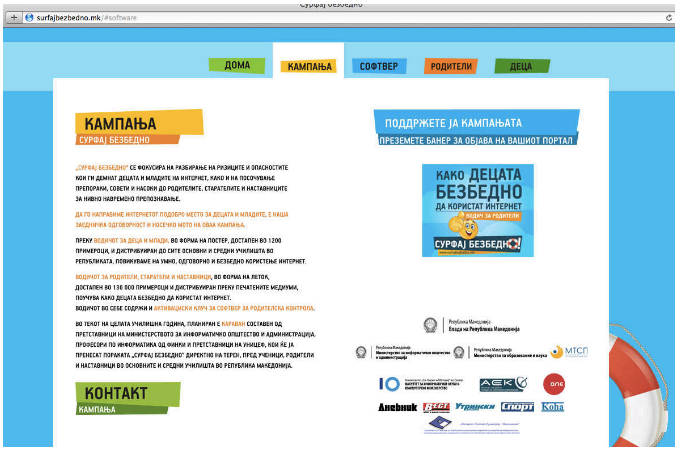
Internet. Target audience were children, their parents, their guardians and their educators. The main online threats that were recognized were:

- Inadvertent exposure to inappropriate images or content
- Solicitation by sexual predators
- Online bullying, harassment or cyberstalking
- Disclosure of personal information.

The industry and the policy makers were tackled by raising the awareness and opening a public debate on the corresponding issues.

II. The Campaign

Within the campaign, there was a web site (Figure 1) set and leaflets and posters printed and distributed. Leaflets (Figure 2) were aimed for the adults, parents, guardians and educators [2]. They were distributed several times, as an insert in the daily newspapers, which were media sponsors of the campaign, and reached about 130.000 x3 adult audience.



„СУРФАЈ БЕЗБЕДНО" СЕ ФОКУСИРА НА РАЗБИРАЊЕ НА РИЗИЦИТЕ И ОПАСНОСТИТЕ КОИ ГИ ДЕМНАТ ДЕЦАТА И МЛАДИТЕ НА ИНТЕРНЕТ, КАКО И НА ПОСОЧУВАЊЕ ПРЕПОРАКИ, СОВЕТИ И НАСОКИ ДО РОДИТЕЛИТЕ, СТАРАТЕЛИТЕ И НАСТАВНИЦИТЕ ЗА НИВНО НАВРЕМЕНО ПРЕПОЗНАВАЊЕ.

ДА ГО НАПРАВИМЕ ИНТЕРНЕТОТ ПОДОБРО МЕСТО ЗА ДЕЦАТА И МЛАДИТЕ, Е НАША ЗАЕДНИЧКА ОДГОВОРНОСТ И НОСЕЧКО МОТО НА ОВАА КАМПАЊА.

ПРЕКУ ВОДИЧОТ ЗА ДЕЦА И МЛАДИ, ВО ФОРМА НА ПОСТЕР, ДОСТАПЕН ВО 1200 ПРИМЕРОЦИ, И ДИСТРИБУИРАН ДО СИТЕ ОСНОВНИ И СРЕДНИ УЧИЛИШТА ВО РЕПУБЛИКАТА, ПОВИКУВАМЕ НА УМНО, ОДГОВОРНО И БЕЗБЕДНО КОРИСТЕЊЕ ИНТЕРНЕТ.

ВОДИЧОТ ЗА РОДИТЕЛИ, СТАРАТЕЛИ И НАСТАВНИЦИ, ВО ФОРМА НА ЛЕТОК, ДОСТАПЕН ВО 130 000 ПРИМЕРОЦИ И ДИСТРИБУИРАН ПРЕКУ ПЕЧАТЕНИТЕ МЕДИУМИ, ПОУЧУВА КАКО ДЕЦАТА БЕЗБЕДНО ДА КОРИСТАТ ИНТЕРНЕТ.
ВОДИЧОТ ВО СЕБЕ СОДРЖИ И АКТИВАЦИСКИ КЛУЧ ЗА СОФТВЕР ЗА РОДИТЕЛСКА КОНТРОЛА.

ВО ТЕКОТ НА ЦЕЛАТА УЧИЛИШНА ГОДИНА, ПЛАНИРАН Е КАРАВАН СОСТАВЕН ОД ПРЕТСТАВНИЦИ НА МИНИСТЕРСТВОТО ЗА ИНФОРМАТИЧКО ОПШТЕСТВО И АДМИНИСТРАЦИЈА, ПРОФЕСОРИ ПО ИНФОРМАТИКА ОД ФИНКИ И ПРЕТСТАВНИЦИ НА УНИЦЕФ, КОИ ЌЕ ЈА ПРЕНЕСАТ ПОРАКАТА „СУРФАЈ БЕЗБЕДНО" ДИРЕКТНО НА ТЕРЕН, ПРЕД УЧЕНИЦИ, РОДИТЕЛИ И НАСТАВНИЦИ ВО ОСНОВНИТЕ И СРЕДНИ УЧИЛИШТА ВО РЕПУБЛИКА МАКЕДОНИЈА.

Fig. 1. Website of the campaign

Posters (Figure 3) were aimed for the children, and were placed inside the schools in the Republic of Macedonia.

Several workshops for the parents and educators in the elementary, middle and high schools were held. One workshop was organized in the Red Cross association, for the guardians of the children in the foster families. Professors of the Faculty of Computer Science and Engineering gave lectures on these workshops, and were moderators of the discussions afterwards. The subject of the lecture was to:

- Identify risks and vulnerabilities to children in cyberspace
- Create awareness
- Develop practical tools to help minimize risk
- Share knowledge and experience when they use Internet and the models of their protection.

As a part of the campaign a free software tool – Internet browser in Macedonian language (Figure 4), for the children in elementary school was distributed to the general public in the Republic of Macedonia.

The Internet browser works as a parental control for the children – it offers the parent to enter a list of approved web sites, a list of banned web sites, and the lists may be updated. Another feature of the Internet browser is that a parent may enter daily or weekly quota for hours spent on Internet, and also a time schedule for the periods of a day that are allowed for Internet use. Another important features are that the browser may be started automatically on log in, desktop, taskbar, minimize and maximize, Alt-Tab, Ctrl-Esc functions may be disabled, and the only way out on Internet for the child is via the controlled Internet browser.

Fig. 2. Leaflet of the campaign

Within the workshops questionnaires were distributed, the results of which will present and discuss further in this paper.

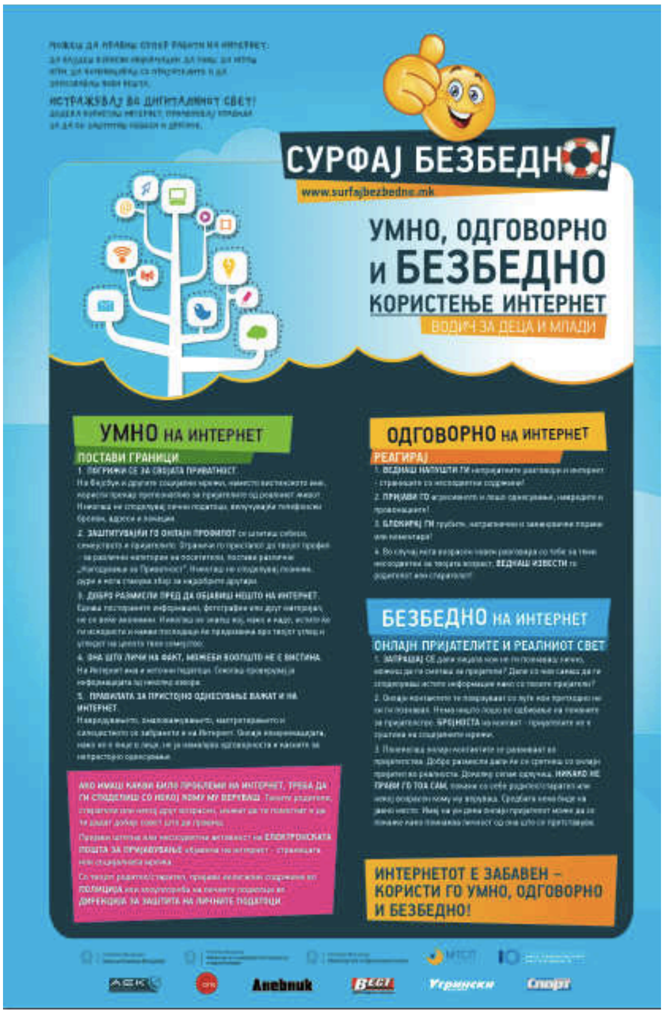


Fig. 3. Poster of the campaign

III. The Survey

In the survey, total of 178 respondents answered the questions stated below. In parentheses after each question is the number of respondents who selected the answers respectively. In some cases, respondents have not completed filling out the questionnaire and some questions had more than one answers selected.

- Education of the parent/guardian

[95] College [11] Two years of college

[43] High school [18] Elementary

- Do you use computer

[159] Yes [8] No

- Do you have a Facebook profile

[126] Yes [39] No

- How old is your child

[101] 5 to 11 [37] 12 to 15 [29] 16 to 18

- Does your child use a computer

[151] Yes [10] No

- How many hours a day he/she spends on the computer

[72] up to 1 hour [68] 1 - 3 hours [22] more than 3h

- Does your child has a Facebook profile

[127] Yes [35] No

- Are you friend on Facebook with your child

[104] Yes [56] No

- Who should be responsible for educating children about their behavior online

[152] parents [40] school [19] government

The last two questions were open ones:

- Can you specify some Internet activities that are punishable by law in the Republic of Macedonia
- What measures should be taken to improve the safety of children in their use of the Internet.

МОЈ ПРЕЛИСТУВАЧ
Интернет прелистувач за деца
• Едноставен интерфејс
• Можност за повеќе корисници
• Лесна навигација
• Различни нивоа на заштита
• Список на дозволени веб содржини
• Систем на заштита преку лозинка
СТУВАЧ

Fig. 4. Internet browser

The respondents had children aged between 5 and 11 years at most (60.5%), which was the main target group of the project as the most vulnerable population age segment. In terms of level of education respondents with higher education dominated (56.9%). This percentage is about five times higher than the percentage of higher educated population in the Republic of Macedonia. Some of the reasons of this educational structure of the respondents are that schools were in urban areas, and parents with higher education were more aware of the threats that arise from Internet, and more interested in taking part in the campaign. Respondents with primary education completed originate mainly from the foster families for the children without parental care, but despite the low standard of living, however the number of families without a computer is less than the number of respondents with primary education (8 vs. 18).

Three of the questions were Facebook related, as Facebook was identified as one of the biggest source of threats in children's' online experience, partly due to very high rate of Facebook users as a percentage of Internet users in Macedonia (82%).

The parents and guardians recognized their responsibility in providing education on children's behavior online (72%).

The answers on the open questions varied, and most common ones were following:

- Can you specify some Internet activities that are punishable by law in the Republic of Macedonia – giving personal information, hate speech, luring, violence, pornography, racism, overtaking another person's profile, entering unauthorised into a web site, publishing other person's photos and information, violent video games, paedophylia, using another person's information, hacking, profile overtaking, personal insults etc.
- What measures should be taken to improve the safety of children in their use of the Internet – secure password, talking to the children and explaining the threats and coping mechanisms, instalation of protective software, school should take place in the efforts to educate children on their safe behavior on Internet, building closenes and mutual trust with the children, organizing of the parents to fight online threats, forming a commitie for online censorship, parents' education, control of a children's online acitivities by their parents, instructions not to contact unknown people, setting proper security settings on Facebook, encoding the web sites with unappropriate content, restricted access on adult web sites, identity protection, cautious communication, fight against game addiction, no Internet – there is no other solution, etc.

IV. Conclusion

The Child online protection campaign conducted in the Republic of Macedonia has met its goals. It raised the awareness regarding online threats for the children, and opened a public debate on the corresponding issues.

The survey demonstrated that the parents, guardians and educators are aware of their role in providing safe Internet experience for the children, but one part of them are not acquainted with the mechanisms how Internet works, which activities are punishable by law in the Republic of Macedonia, and what are the proper measures to improve the safety of children in their use of the Internet.

The conclusion arising from the campaign is that the education of children, their parents, guardians and educators should be an ongoing process, system approach should be implemented, and apart from the campaigns, it should become a part of the traditional education process.

References

[1] http://www.itu.int/osg/csd/cybersecurity/gca/cop/
[2] http://surfajbezbedno.mk
[3] http://www.onguardonline.gov/topics/protect-kids-online
[4] https://www.getsafeonline.org/safeguarding-children/
[5] http://www.microsoft.com/security/family-safety/childsafety-age.aspx
[6] http://home.mcafee.com/advicecenter/?id=ad_fis_htpco&ctst=1
[7] http://www.us-cert.gov/ncas/tips/ST05-002

Current state of E-Learning with an Emphasis in the Balkan Region

Blerta Abazi Caushi
Faculty of Business and Economics
South East European University
Tetovo, Macedonia
b.abazi@seeu.edu.mk

Agron Caushi
Faculty of Business and Economics
South East European University
Tetovo, Macedonia
a.caushi@seeu.edu.mk

Abstract— **Higher education is not being refrained from the technological progress. To adapt to the changes, institutions of HE have to integrate ICT solutions, and especially e-learning in their courses and programs. This paper provides a discussion of the influence of information technology on higher education and a review of the literature on e-learning evolution and environments. Moreover, in this paper we use two frameworks for e-learning to distinguish and recommend suggestions for adoption and improvement of the e-learning systems within universities. Key trends and statistics regarding the e-learning usage in the Balkan region are identified through a survey conducted in 40 universities. Our results contribute to the understanding of the current level of adoption of these systems and the reduction of the gap that exists in research for the e-learning environments and usage in the Balkans.**

Keywords: ICT, HE, e-learning, online learning, distance education, information systems

I. INTRODUCTION

There is no doubt that we are living in a very technology challenged setting. Whether that is in our working places, at home, or at our learning environments – we are surrounded by a certain technology in use. Smartphone's, tablets, laptops on one side, fast internet access, huge amount of available data, great end-user connection on the other side – make us think of the way educational institutions have to change to adapt to this scene. In this era of technological society, there is no other way than to approach these new tech savvy students with planned ICT solutions. It is more than evident that within the institutions of higher education, there is constant innovation in terms of provisioning of online services as a response to new challenges. Learning Management Systems, distance education and online learning, have become an important feature of online service delivery within the Higher Education Information Services sector, requiring close attention to issues of functionality, sustainability and usability.

This paper discusses the impact of technology in University setting, and examines the importance of e-learning environments. We go through an analysis of the phases of evolution of e-learning and also discuss about the e-learning environments. We pinpoint the importance of adding this important element into the strategic plans of the universities, especially after using two e-learning frameworks. We find that these frameworks are inter-related and show that they can be used in already existing e-learning environments. The result is seven recommendations that can be adopted by these institutions to better integrate their e-learning systems with their strategic plans and operations. Moreover, we examine the application of the types of e-learning system in Balkan countries. In our research we survey 40 Universities in 10 countries in this region, and we find the levels of e-learning adoption and the systems they use. Finally, our analysis support the idea that e-learning should not be viewed only as technology solution; rather it should be embedded in the strategic plans of the institutions to utilize its benefit and potential [1].

II. E-LEARNING SHAPE IN HIGHER EDUCATION INSTITUTIONS

The changing landscape of HEIs in order to achieve successful studying experience and sustainable development is being witnessed through continuous alteration in the approaches to new technology. E-learning has had a big impact into institution's overall strategy, and they have implemented it with the aim to improve the technology for education and increase the quality of the educational process. According to this and responding to the challenges of the 21st century, HEIs saw knowledge and ability to use e - learning as an integral part of basic literacy to every member of the academic community and considered as one of the ways for active involvement in the University of continuous lifelong learning of students. We believe that in today's university landscape, a key strategic issue regarding online learning is not whether to engage, but how [1]. The first step towards this commitment is to create or buy/rent learning management system (LMS). While there are several definitions of LMS, we chose the one by Ryann K. Ellis [2]: "a software application that automates the administration, tracking, and reporting of training events." According to the same author some of the characteristics that a robust LMS should have are: centralize and automate administration, use self-service and self-guided services, assemble and deliver learning content rapidly, consolidate training initiatives on a scalable web-based platform, support portability and standards, personalize content and enable knowledge reuse [3].

A. Evolution and environments

A report from Educause, that surveys more than 2000 universities in the world each year from year 2000, shows that the issue of e-learning environments has been among the top ten issues for these institutions throughout the period 2000-2014 [1]. To be able to understand why such an importance is given to e-learning environments, and especially learning management systems (LMS), we need to take a loop back to see are the features of e-learning, what is learning management system and how did it evolve through years. To avoid misunderstandings, in our paper, which is in accordance with several studies, we define the online courses, those courses in which at least 80 percent of the course content is delivered online [4]. Many authors use different terminology for certain types of service delivery through ICT. First, we have the term *distance education*, which generally means creation of online environment to support students that are geographically far away, and impossible for them to attend lectures in the traditional university setting. Second, the term *online learning* usually refers to allowing students get their degree completely online through virtual universities and so on. And third, *e-learning* is defined as a method of delivering a course online, for students who also are enrolled in university and attend courses the regular way. A study conducted by Moore, Deane and Galyen [5] tries to make a comparison of the three terms. However, their finding show that perhaps in the past years it made a distinction, but now they are used indivisible manner. In our study, we find that HEIs do use all the three terminologies ambiguously, and they in fact are used to depict different situations, but essentially they refer to providing teaching and lecturing in an online manner. What we want to pinpoint in this paper is that no matter how we call the online service delivery, they all have one thing in common: They need a system in place to allow delivering of the content rapidly, automate administration, personalize content and enable knowledge reuse [2].These systems broadly are called learning management systems, and the next section focuses on the development of these systems that are of crucial importance in today's university setting.

The earliest Learning Management System (LMS) can be dated back to Sydney's Automatic Teacher in 1924, a primitive system that automated grading of multiple choice tests [6]. However, this was the first towards what we call today Learning Management Systems. In 1956 SAKI was invented [7], a system that would automatically adjust the difficulty of questions based on the performance of the user. 1969 Arpanet, the precursor of today's web was created, which will have a huge impact in the way LMS will develop. It was until 1997 that interactive learning network was designed. Courseinfo is amongst the first players in this sphere [8]. In year 2002 Moodle was released, which is today one of the most used LMS in university settings in the world. And today, most modern LMS are hosted in the cloud, which make much easier the process of moving to e-learning environment since there is no need to install or maintain a system and especially no need for the burden of in-house development [9].

During the beginning of this millennium, most universities in the world had started to develop asynchronous learning environments and services to support students involved outside of the campus. The early 2000's were the beginning of delivering education outside of the classroom. Over the next few years, we see a slight shift from the traditional classrooms to more network based teaching, and the beginning of the transformation of the institutions [10]. The efforts are mainly in creating hybrid learning experiences for the students both on-campus and out. It is in the year 2005 that these programs become available at colleges and universities in different formats, as certificates, diplomas, degrees, or post-baccalaureate programs. From year 2006 E-learning comes to light from its initial form as an "an add-on to traditional education" to "mission-critical component of the educational environment" [11]. The few last years, are the years of the rapid development of technology-based learning tools. According to a study [12]: "This rise in strategic importance is evidence that technology has moved beyond the data center and institutional administrative systems and is now part of daily life for faculty and students".

Consistent with a research conducted by Elaine Allen and Jeff Seaman based on responses from over 2,800 Chief Academic Officers (CAOs) and academic leaders, 69.1% of the respondents agreed that online learning is of strategic importance for the university [4].

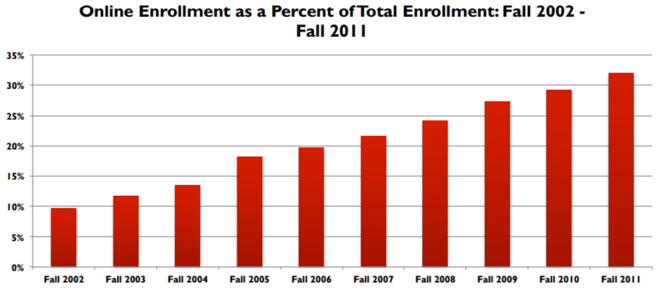


Figure 1 Increase in Online Enrollment in USA

The reason for its strategic importance is that the number of the students that take online courses has increased a lot in this 10 year period (see Fugure 1). It is a fact that around 72% of universities had online offerings even ten years ago. A major change that is worth mentioning is that a far larger proportion of higher education institutions have moved from offering only online courses to providing complete online programs (62.4% in 2012 as compared to 34.5% in 2002) [4]. The next section discusses the e-learning frameworks that help HEI adopt easier the challenges of e-learning.

B. E-Learning Framework

Faced with the need of transforming the university structure, processes, and programs according to the Bologna reform, and in order to become more flexible and more responsive to the environment, e-learning implementation takes the role of an **institutional** and **strategically planned operation** [13]. E-learning frameworks aim to improve the reusability of design and implementation by following the practices originated in software engineering [14]. There are several frameworks in place to support institutions of HE to approach e-learning. In this paper we discuss only two most commonly used frameworks and we conclude the section with seven recommendations that can be adopted by these institutions to better integrate their e-learning systems with their strategic plans and operations.

The 8 dimensional e-learning framework [15] is one of the most prominent frameworks used by scholars worldwide. These 8 dimensions provide an in-depth analysis of the issues that arise with the implementation of the e-learning, especially into teaching and learning experience for both the instructors as well as the students.(see figure 2) This model captures the core functionalities starting from the issues of content, media, design analysis in the pedagogical dimension, to the issues of infrastructure support, usability, assessment of the users, maintenance and reusability, learner diversity and legal issues, and ends with the administrative and academic affairs as well as with student service issues in the institutional dimension.

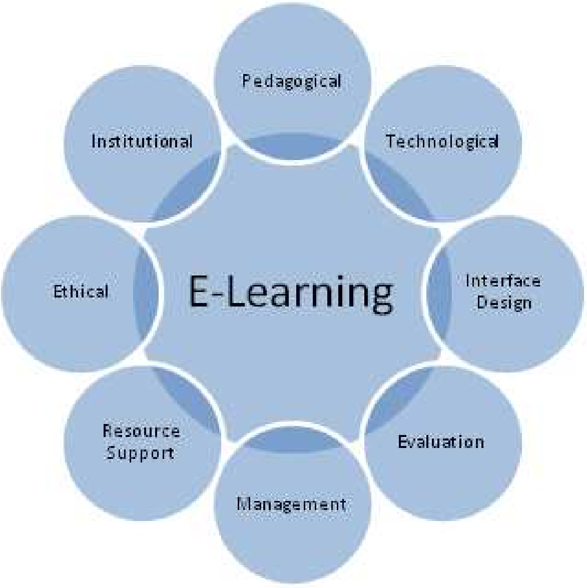


Figure 2 The 8 Dimensional framework for e-learning

Universities using this framework, can grasp every issue in detail, integrate them and create constraints - so that the final product captures all the issues that rise with this framework.

The other is the *TPACK framework.* There is a lot of resemblance with the 8 dimensional model, since both these frameworks have in their heart the forms of knowledge that can be engendered: technological, content and pedagogical knowledge. This framework was chosen because of its similarity with the 8 dimensional model so that it positions the importance of LMS in the strategic plans of the institutions, due to the many stakeholders and many issues that arise not only from the IT perspective. Using this framework, as you can see from the figure 3 [16], institutions of HE can achieve effective integration between the three different types of knowledge generation, allowing for development of sensitivity to the dynamic relationship amongst the different components in a unique context. The end result is the creation and maintenance of TPACK (technological pedagogical content knowledge) which uniquely will adapt to the specific needs of the different courses, with different student background, different difficulty levels, and more importantly with diverse technological skills for both the instructor and the user, i.e. the students.

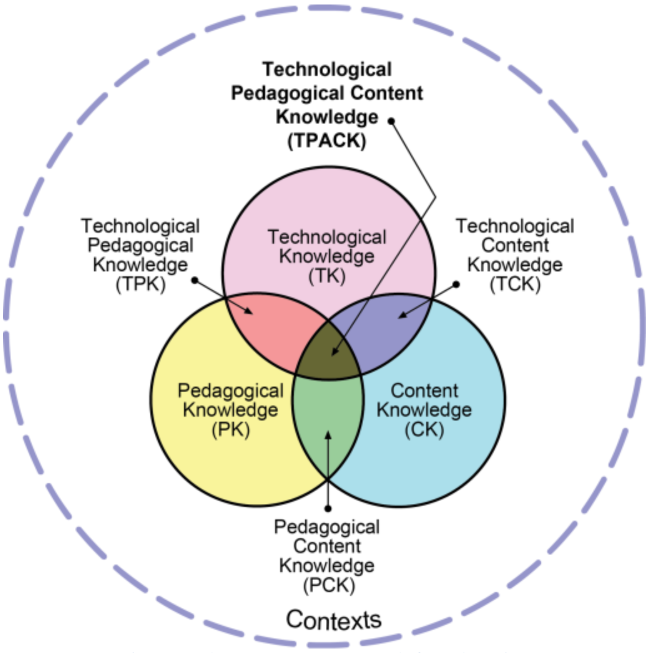

Technological Pedagogical Content Knowledge (TPACK)
Technological Pedagogical Knowledge (TPK)
Technological Knowledge (TK)
Technological Content Knowledge (TCK)
Pedagogical Knowledge (PK)
Content Knowledge (CK)
Pedagogical Content Knowledge (PCK)
Contexts

Figure 3 The TPACK Framework for e-learning

What is evident from these two frameworks is that they both raise the issue of e-learning as a strategic initiative and operation of the institution that takes this endeavor. Several advices for universities that have adopted or plan to adopt e-learning, and especially for the institutions that want to achieve operational effectiveness as well as increase of quality in terms of e-learning usage can be drawn. Based on different studies and research done in this sphere, as well as in line with these two frameworks, we conclude the following suggestions for improvement:

1. Institutions of Higher Education should make e-learning initiatives part of the institution's strategic plan and budget, and set specific goals for e-learning initiatives.

2. HEIs should try to centralize essential e-learning technology services as much as possible because of the greater efficiency and seamless integration of e-learning services.

3. E-learning should be viewed as critical to the mission of the university and the provision of e-learning services should have high priority within IT.

4. The reliability of the technology used in e-learning should be seamless.

5. The university should create a clear path that will demonstrate the merits of e-learning for *both* traditional face-to-face and online classrooms. It is important to keep the faculty and students interested in this new learning environment, and it is a good add-on to implementing a faculty e-learning mentoring program using faculty who have already taught e-learning courses.

6. Universities should ensure that the chosen technologies for e-learning are scalable, by creating a plan for the number of courses/programs that will roll out e-learning initiatives in the coming years

7. Universities should make sure that the chosen technologies adaptable – not all the courses and programs have a need for the same technology.

The next section discusses briefly the e-learning trends in the world, and focuses more in the adoption level of e-learning in the universities in the western Balkan countries.

III. THE CURRENT STATE OF E-LEARNING IN THE BALKAN COUNTRIES

Recent developments in HEI, according to latest CDS survey, shows that institutions that have already adopted LMS in the cloud reaches 12% of the respondents. Another 65% have online learning platform in place, and only 9% of institutions do not have discussion to date about online learning [17]. In this section we analyze the current state of e-learning in the western Balkan region and try to contribute to better understanding of the setting that our universities operate here.

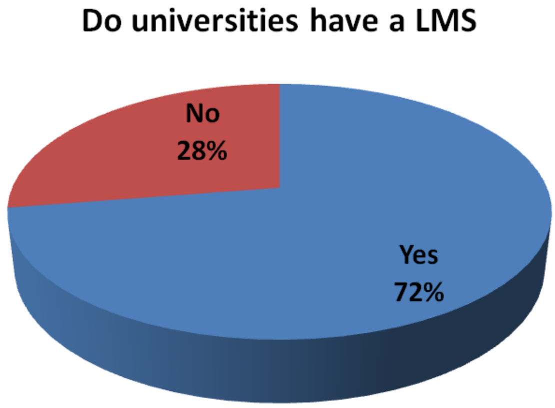


Figure 4: LMS adoption level in Universities of Balkan Region

For the purpose of this paper we analyzed 40 universities from 10 countries in Balkan region including Macedonia, Albania, Greece, Bulgaria, Serbia, Kosovo, Montenegro, Bosnia and Herzegovina, Croatia and Slovenia. We tried to answer the following questions: Does the university use any learning management systems; what type of system do they use, and what specific product do they use. The data was acquired by conducting a short survey that covered only the abovementioned issues. We also visited their websites, and in cases there was not enough information, we would contact representatives of the university by asking questions that we were interested in. We are aware that the data might not be 100% accurate and that we have not included all the universities, yet we think the results of this study are important and reflect the reality of current state of e-learning in the region. These figures are just a sample, and data might vary if all the universities were surveyed.'

According to our study, about 70% of the universities in this region have some system in place for uploading course content as lecture notes, assignments, etc. From the Figure 5, we can see that Greece, Bulgaria, Croatia and Montenegro have a LMS in place in all of the surveyed universities, whereas in the rest of the countries we could still find universities that are not using any LMS as a part of regular or distance learning studies.

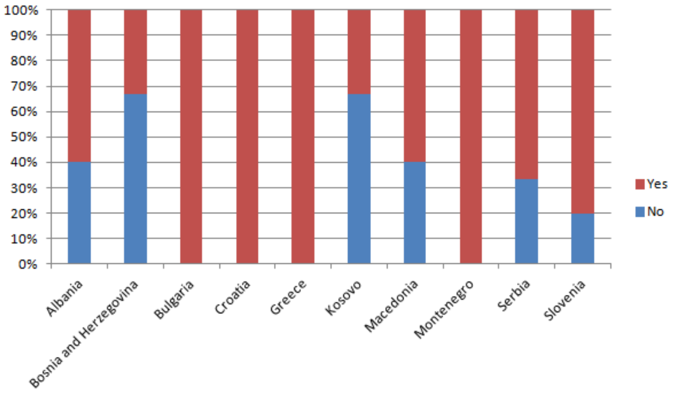


Figure 5: LMS Adoption Level by Country

The second question that this study was trying to find an answer to was to find out what type of LMS do universities use (see Figure 6). The results show that, of the universities that have a Learning Management System in place, 43% have created their LMS in-house whereas the others have implemented a readily available package. From the LMS packages, Moodle is definitely the most popular one. We assume that the main determinants for choosing Moodle are that it is free, it is open-source and it has a big community.

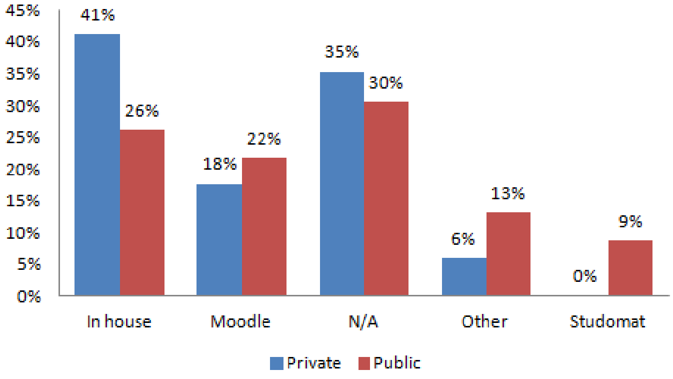


Figure 6: Types of LMS adopted in Public and Private Universities

Also during this study, we analyzed if there is a difference in implementation of LMS depending on the type of the university. In our study 17 universities were private and 23 universities were public. Based on the data that we collected, there is no significant difference between and private universities on implementation of LMS. As we can see from Figure 7, around 77% of private universities and 70% of the public universities have implemented a LMS. Interestingly, we can see a similarity in the preference of a LMS package, with some minor differences. The major difference can be seen in the preference of choosing a readily available solution or building the LMS in-house. 41% of the private universities have chosen to build their own LMS versus 26% of the public universities with the same choice. It is worth to mention that all the analyzed public universities in Croatia use and LMS called Studomat.

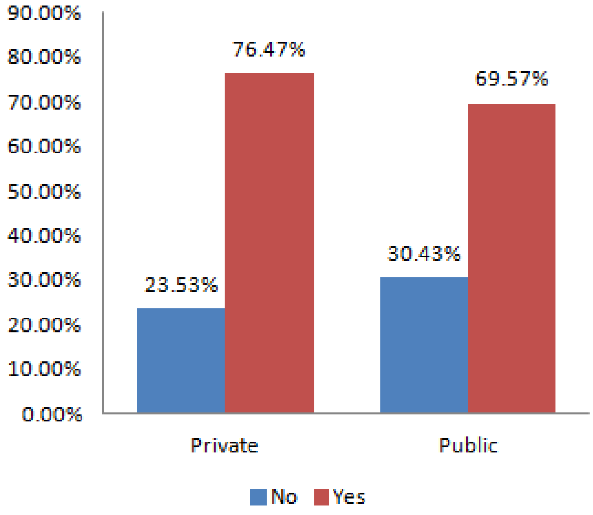


Figure 7: LMS Adoption Level in Public and Private Universities

In general, the results of this study show that, Balkan region is not lacking in implementing Learning Management Systems as a technology or platform for e-learning. The tendency is to develop a LMS in-house or adopt an open source or free LMS which is understandable due to lack of funding for IT. However, this study is lacking answer to the question, how much are these LMSs used in reality and to what extent do they fulfill their mission of transferring knowledge through online media. These questions need a broader research where many universities will be involved and a better qualitative and quantitative data will be collected to answer the questions that are addressed but not answered as a part of this research.

There were several initiatives to improve and enhance the quality of e-learning in the Balkan region. One of the project that just ended last year is the DL@WEB project has been designed to improve the quality and relevance of distance education (DL) at Western Balkan higher education institutions and to enable easier inclusion of partner country institutions into European Higher Education Area. There were 6 Universities from this region (Serbia-3, Montenegro-2, and Macedonia-1) working with three universities from European Union countries (Italy, France and Slovenia). The outcomes of this specific project, development and implementation of accreditation standards, guidelines and procedures for quality assurance of DL study programs at national system levels in WB countries are not published yet, although the project is over.

The other initiative similar to the previous project is the NeReLa (network of remote labs) project, that aims, amongst other things, to introduce innovative teaching methods. This project is a joint effort of Spain, Portugal, Cyprus, Slovenia and Greece from EU countries, and Serbia where beside national institutions and associations, also 4 universities are members of this endeavor.

The purpose of these projects and so many other ongoing projects in this region as the FETCH project for “How to support learning at anytime anywhere” which is due year 2020, are to improve distance education and e-learning experience within institutions of HEI. Due to the Bologna Declaration (19.06.1999) with the joint declaration from the European Ministers of Education, the major point for development is the enhancement of quality of teaching within institutions of HE in Europe. According to this declaration, the main focus is the acceleration of student, graduates and HE staff mobility, as well as access to high-quality HE. In line with this, universities in Balkan region have to comply with the trends and they need to work in terms of defining what constitutes high quality education. E-learning is definitely a big part of it, especially since it takes such a great role in strategic plans of universities throughout the world.

IV. CONCLUSION

In summary, there is a growing recognition of the need for e-learning solutions in the Balkan region. Institutions are becoming aware of the trends in e-learning. As we saw earlier, most of the universities are making their steps in adopting Learning Management Systems and some of the universities have started with offering some distance learning and online programs. From our analysis, we notice that all these universities that were part of the short survey, either had open source, either build in-house LMS. We believe that none of these universities has gone for a vendor based solution because they are in the beginning stages of e-learning and they do not want to spend funds for a vendor based solutions. However, it is encouraging to see that both, public and private universities embrace the trends in higher education. They have started to understand that new generations are technology native and they need things anywhere, anytime. In spite of this, we should highlight that e-learning is not only a technological matter. Rather, it is a process involving academic staff, students, and pedagogical content. For this reason, it is important for universities to have a strategic approach to e-learning. It takes a lot more than providing a technological platform for e-learning in order to be successful in transferring knowledge to students using electronic media. If higher education institutions view e-learning purely as a technology, they will be doomed to fail [18]. According to a study from Educause regarding e-learning: “*the issue of determining the role for online learning and developing a strategy for that role is under constant revision by faculty, administrators, instructional designers, and, most important, the IT personnel and resources that make it all happen. Clearly, the amorphous nature of online learning requires agile and adaptable strategies, along with strategists who are committed to furthering student learning and to* *finding the best means to accomplish this singular goal that is at the heart of all institutional activity*" [1].

Acknowledgment

We would like to express our special thanks of gratitude to our student in the Information Systems and Management program, Milka Trajkova (mt19992@seeu.edu.mk) for her great work in the process of data collection and surveying universities regarding e-learning systems.

References

[1] S. Grajek, "Top-Ten IT Issues, 2014," EduCause, 2014.

[2] R. Ellis, "Field Guide to Learning Management Systems," ASTD, 2009.

[3] A. A. Sejzi and B. Arisa, "Learning Management System (LMS) and Learning Content Management System (LCMS) at Virtual University."

[4] I. E. Allen and J. Seaman, Changing Course: Ten Years of Tracking Online Education in the United States. ERIC, 2013.

[5] J. L. Moore, C. Dickson-Deane, and K. Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," Internet High. Educ., vol. 14, no. 2, pp. 129–135, Mar. 2011.

[6] L. T. Benjamin, "A history of teaching machines.," Am. Psychol., vol. 43, no. 9, pp. 703–712, 1988.

[7] G. Pask, "SAKI: Twenty-five years of adaptive training into the microprocessor era," Int. J. Man-Mach. Stud., vol. 17, no. 1, pp. 69–74, Jul. 1982.

[8] D. M. Fahey, "Blackboard COURSEINFO: Supplementing In-Class Teaching with the Internet," Hist. Comput. Rev., vol. 16, no. 1, pp. 29–37, Jan. 2000.

[9] F. M. M. Neto and F. V. Brasileiro, Eds., Advances in Computer-Supported Learning: IGI Global, 2006.

[10] B. Abazi Caushi, A. Caushi, and Z. Dika, "A Comprehensive Aproach to Technology Issues and Challenges for Higher Education Institutions," in 12th International conference e-Society, Madrid, Spain, 2013, pp. 177–184.

[11] B. I. Dewey and P. B. DeBlois, "Top-ten IT issues, 2006," Educ. Rev., vol. 41, no. 3, p. 58, 2006.

[12] B. L. Ingerman and C. Yang, "Top-Ten Issues, 2011."

[13] M. Zuvic-Butorac, Z. Nebic, D. Nemcanin, T. Mikac, and P. Lucin, "Establishing an Institutional Framework for an E-learning Implementation–Experiences from the University of Rijeka, Croatia.," J. Inf. Technol. Educ., vol. 10, 2011.

[14] M. Derntl and R. A. Calvo, "E-learning frameworks: facilitating the implementation of educational design patterns," Int. J. Technol. Enhanc. Learn., vol. 3, no. 3, pp. 284–296, 2011.

[15] B. H. Khan, Managing e-learning: design, delivery, implementation, and evaluation. Hershey, PA: Information Science Pub., 2005.

[16] M. Koehler, P. Mishra, M. Koehler, and P. Mishra, "What is Technological Pedagogical Content Knowledge (TPACK)?," Contemp. Issues Technol. Teach. Educ., vol. 9, no. 1, pp. 60–70, 2009.

[17] L. Lang, "CDS Executive Summary Report," 2013.

[18] G. Ssekakubo, H. Suleman, and G. Marsden, "Issues of Adoption: Have e-Learning Management Systems Fulfilled Their Potential in Developing Countries?," in Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference on Knowledge, Innovation and Leadership in a Diverse, Multidisciplinary Environment, New York, NY, USA, 2011, pp. 231–238.

E-system for 360 Degree Instructor's Evaluation

Veno Pachovski, Zlatko Ivanovski, Eva Blazevska
School of Computer Science and Information Technology
UACS
Skopje, Macedonia
{ pachovski, zivanovski, blazevska} @uacs.edu.mk

Ana Krleska
School of Business Economics and Management
UACS
Skopje, Macedonia
ana.krleska@uacs.edu.mk

***Abstract* — The evolution of the contemporary higher education means perpetual change, upgrade and evolution – improving the teaching methods as well as evaluating the level of acquired knowledge. There is a need for constant monitoring of the teaching process – and how the students perceive it. Therefore, there must be a way to follow the level of student satisfaction since students are direct participants in the process (whether it is connected to the teaching or the functioning of the university in general), as well as monitoring the performance of the instructors. In order to gets a holistic and objective input on the engagement of the instructors; the 3600 evaluation is a highly recommended tool for performance evaluation. The 360^0 evaluation [at University American College Skopje] contains several elements including but not limited to:**

1. **self-evaluation,**
2. **peer-to-peer evaluation,**
3. **evaluation by dean,**
4. **evaluation by Chief Academic Officer**
5. **evaluation by administration,**
6. **evaluation by the Rector and**
7. **Student evaluation**

For the purposes of this paper, we will be discussing the last parameter – the student evaluation and in particular, the e-system for electronic student evaluation. The model offers secure and anonymous data collection, automatic data processing and various reports in order to present the results. It also describes the appropriate methodology applied for data analysis that displays the results of the instructor for each course, the performance for that semester and/or academic year, the whole department, results of all instructors teaching at one school or the overall performance of the instructors at the university level. Based on the proposed model, e-system is developed and is being successfully implemented at a private university in the Republic of Macedonia for the last five years. (*Abstract*)

***Keywords*— education, evaluation, 360 degree, performance evaluation, statistics**

I. INTRODUCTION

The tool for 360° employee evaluation is a commonly used by the human resource departments in companies and organizations worldwide as part of the performance management systems. It is considered a highly effective evaluation tool for enhancing employee development and professional growth (Edwards & Ewen, 1996). It is a mechanism for evaluating someone's performance based on feedback from all parties concerned by the individual's work — supervisors, co-workers and subordinates (within the company) as well as, partners, customers, or the general public as externally affected parties. It is a method of collecting input from various sources of the employee's environment [1].

This model was primarily used to gather employee's opinion about important questions [2]. The model was later developed and registered in the mid-1980s by Teams, Inc., and has been successfully implemented by over 90% of Fortune 500 companies. The 360° evaluation model seeks to provide feedback on employee's strengths and weaknesses using evaluation reports from their supervisors, peers, clients, and other parties with whom they interact in the conduct of their jobs. The feedback has not been used to determine salary adjustments so far in any of the organizations that have implemented it [3].

In the model that we used, we assume that the instructors' self-report combined with evaluation reports by the administration, peers, dean, Rector, Chief Academic Officer and students are the best tool for performance measurement in education.. The aggregate report provides the basis for dialogue between the instructors and their supervisors (dean, rector, Chief Academic Officer etc.) to arrive at a final evaluation rating and to improve the instructor's performance and attitude wherever is deemed necessary [4].

A. Motivation and problem approach

One of the best methods for diagnosing issues in education is implementationof questionnaires. This is one of the fastest ways to reach students and get their feedback for important matters, especially if the institution is interested in getting direct feedback in terms of the teaching process. The evaluation is conducted anonymously and students are free to express their personal views and impressions for each course they take using a questionnaire designed specifically for this need. The questionnaire encompasses the following categories:

- Syllabus and course materials
- Course delivery
- Instructor's Assessment

The questionnaires provide important results which are used to improve the overall education process. The setback is that the student evaluation is usually conducted by using traditional pen and paper questionnaires which requires sufficient amount of resources and it is a time consuming process. For a university that strives to protect the environment, using large amounts of non-reusable paper is not the most desirable practice. In addition, once the questionnaires are filled-in, the time consuming process of data processing begins in order to make sense of the data gathered. This assumes that each answer of each questionnaire should be inserted individually in some analytical program that will process the raw data. It takes approximately one minute to insert each questionnaire into the system. For two semesters, for five courses each semester, for six schools for an average of 20 students per group and for an average of 2 instructors per course… calculations are not in favor of time as a factor. To add up, human error during the entire process is evidently inevitable.

In spite the benefits in terms of cost reduction, the most important benefit that an electronic system could provide is the anonymity and individuality of the entire process. With pen and paper, students are often reluctant to give straight forward answers on questions, especially when the grade should be negative. Instead, they are more prone to giving higher grades believing that if the anonymity is breached, at least they could earn 'positive points' with the professor by giving them a high score.

With the development of the IT technologies, especially the web applications, many universities are willing to abandon the traditional pen and paper evaluation process and to implement electronic evaluation systems. The initial implementation of the system might be expensive, but on the long run it returns the investment and cuts the future expenditures.

The goal of this paper is to present the model which is used for the development of the web application used for student evaluation and data analysis. The application provides: fast and efficient evaluation, anonymity, minimal cost, less time-consuming automated data processing and reports, data archive for comparative reports, quantitative and qualitative data and much more. At the same time, the application is protected with several levels of access, enabling different user account types. And finally, the system minimizes human error as a factor in the process of data analysis and data processing.

II. TECHNICAL DESCRIPTION OF THE SYSTEM

The model contains modules for three different types or users (actors).

The first actor is the administrator of the application. The administrator is responsible for initial setup of the entire system including: the structuring of the questionnaires, defining the basic nomenclatures (classifications) of the system and creating the links between the different categories, coordinating the process in terms of activation, deactivation and re-activation, and finally summarizing the results in useful and comprehensible reports. The second group of actors consists of the instructors who are given the opportunity to access the evaluation program and to activate the evaluation which is predefined by the administrator and is intended for their course and/or their group of students. The model offers instructors to fill-in data in a self-survey for each course they teach. The instructor can fill-in survey for his/her peers as well i.e. people that work in same department/school.

The third group of actors in the system are the students who do not have direct access to the structure of the application. They are only allowed to fill-in the data in the electronic survey form for every teacher-course combination they participate in. The students are provided with anonymity in order to obtain the most realistic and objective results for student satisfaction for each question of the questionnaire.

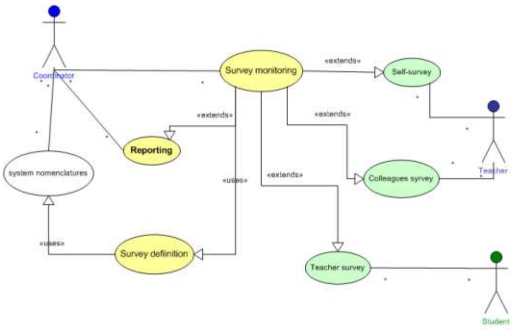


Fig. 1. System architecture – roles and activity flows

A. *Underlying database structure*

In order to provide more reliable environment for conducting the evaluation process, the system uses relational database. The core of the database used for the evaluation process consists of 10 charts. Basic information about every evaluation survey is kept in chart anketa_hd. This chart is automatically filled with data whenever the coordinator starts the process of evaluation. Previously, this user must create all the connections between the nomenclatures (teachers, courses, questions for specific survey etc.) so that the system is able to generate the exact teacher-course surveys. Every record of this chart contains information about the teacher and course the survey is about, its date of creation and activation.

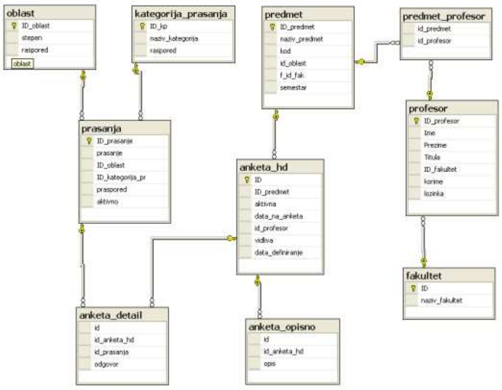


Fig. 2. System's database architecture – charts and relationships

The data gathered when a user (student) fills-in the appropriate questionnaire are kept in anketa_detail and anketa_opisno charts, both presented in Fig 2. These charts are the most important ones in the reporting process because their records are the ones that contain the results that are processed according the given methodology. Every record contains information about the question answered, the survey it belongs to and the answer given by the respondent user.

B. Evaluation algorithm

The evaluation process consists of several steps that must be performed by every actor that uses the system. Coordinator has the main role because this user is in charge of the process of creation and activation and monitors the activities of other users.

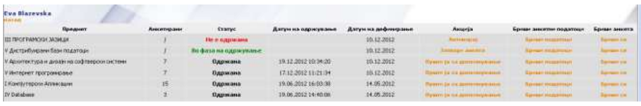
At first, the coordinator must define all the teachers and courses that will be part of the evaluation process. This activity is done at the beginning of the system implementation and it is repeated on demand.

Then, before creation of the evaluation surveys, the coordinator must create correct links between the teachers and courses i.e. it has to define which teacher is in charge for which course. The relation is many-to-many.

When creating a survey, the coordinator has to enter the questions that are part of it. The system allows the coordinator to insert, categorize, modify, activate and deactivate the questions from the surveys. Every question can be categorized by means of: type of survey, survey segment, department etc.

After all these activities are done, the coordinator can generate the evaluation pool for all existing combinations of teacher-course pairs, or only for some of them. The coordinator simply chooses the date from when the evaluation is active and the type of surveys that should be activated. The coordinator has the ability to redefine the evaluation for some date, to make overview of the previously defined evaluations or to delete the blank/unused evaluations defined on a specific date.

When the evaluations are created, either coordinator or every teacher individually activates the surveys. The monitoring of the whole activity is done by several stages the survey can reach during its lifecycle: Active, on Hold, Reactivated, Done or Deleted. The coordinator can easily change the state of the survey.

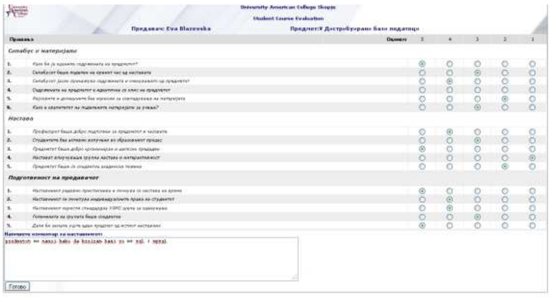


Eva Blazevska

Предмет	Анкетирани	Статус	Датум на одржување	Датум на дефинирање	Акција	Бришење анкетни податоци	Бришење анкета
III ПРОГРАМСКИ ЈАЗИЦИ	/	Не е одржана		10.12.2012	[illegible]	[illegible]	[illegible]
V Дистрибуирани бази податоци	/	Во фаза на одржување		10.12.2012	[illegible]	[illegible]	[illegible]
V Архитектура и дизајн на софтверски системи	7	Одржана	19.12.2012 10:34:20	10.12.2012	[illegible]	[illegible]	[illegible]
V Интернет програмирање	7	Одржана	17.12.2012 11:21:34	10.12.2012	[illegible]	[illegible]	[illegible]
I Компјутерски Апликации	15	Одржана	19.06.2012 16:03:38	14.05.2012	[illegible]	[illegible]	[illegible]
IV Database	3	Одржана	19.06.2012 14:40:08	14.05.2012	[illegible]	[illegible]	[illegible]

Fig. 3. Surveys lifecycle – example of states of some surveys

Once the survey is activated, other users can fill-in the questionnaires. Every user has list of available surveys. The user can choose survey for answering. Every survey can be answered only once by one user. This activity does not require login to the system.



University American College Skopje
Student Course Evaluation
Предавач: Eva Blazevska Предмет:V Дистрибуирани бази податоци

Силабус и материјали

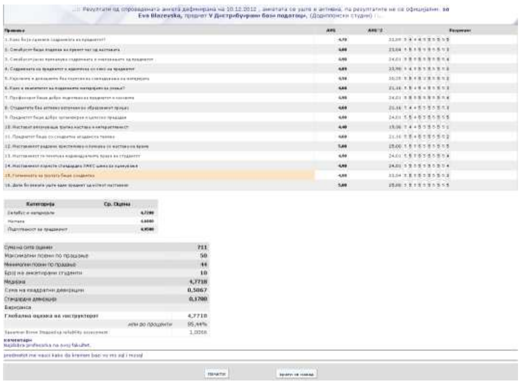
Настава

Подготвеност на предавачот

Готово

Fig. 4. A sample of a questionnaire –screenshot

After the evaluation process for a particular teacher-course survey is done, the survey is closed (state:done) so the coordinator can continue with the data processing i.e. generating reports.



Eva Blazevska, предмет V Дистрибуирани бази податоци

Категорија	Ср. Оценка
Силабус и материјали	[illegible]
Настава	[illegible]
Подготвеност на предавачот	[illegible]

Сума на сите оценки	711
Максимални поени по прашање	50
Минимални поени по прашање	44
Број на анкетирани студенти	10
Медијана	4,7718
Сума на квадратни девијации	0,5067
Стандардна девијација	0,1780
Варијанса	
Глобална оценка на инструкторот	4,7718
или во проценти	95,44%
Spearman Brown Stepped up reliability assessment	1,0066

Fig. 5. A sample of a report – presenting aggregated data

C. Reporting

The system enables several types of reports typical for the 360 degree evaluation. All reports are relevant after the defined period (date) for that evaluation cycle is closed. Otherwise, all reports are marked as "unofficial results".

The most important reports the system generates are:

Students' evaluation for teacher – course. This report contains important details about the performance of the instructor on a particular course. At the top of the report we can find the date and time when the evaluation was conducted. Bellow we have a chart which contains a list of all questions and their corresponding individual grades, the average grade per question and the overall grade of the instructor for that course. . Because all questions are divided in three categories (see p.1 of this paper), there is another chart containing average grades per category.

The summative results are listed in a new chart. Here, there is the following information: sum of all grades, maximum and minimum number of points per question, the total number of students that have completed the evaluation, median, standard deviation, sum of squared deviations, variance and a Spearman Brown Stepped Up reliability assessment. The report also contains a list of anonymous qualitative comments from the students. When the report is printed, a graphic representation of all grades is included, as well.

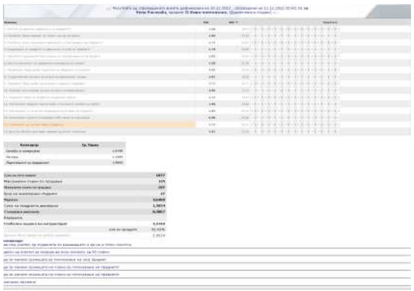


Fig. 6.a. Teacher/course report – grading

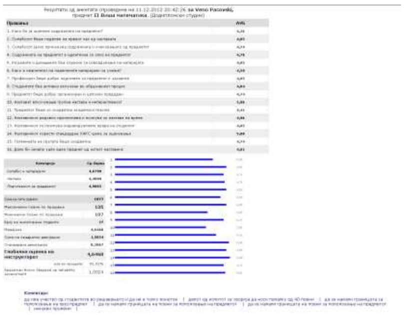
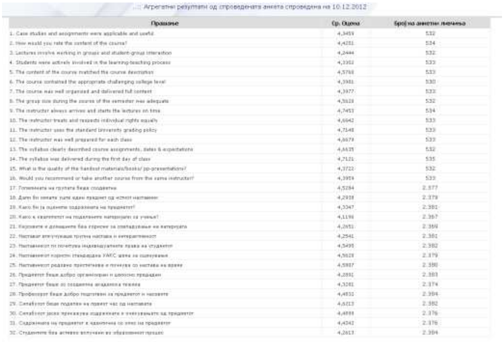


Fig. 7.b. Teacher/course report – graphic representation

Student global satisfaction based on some of the questions in the survey. This report presents the participant's level of satisfaction for particular area. The coordinator of the evaluation process has an option to select whether to include answers from all participants from the whole university or to include participants from desired school. At the end the result from this report can be used to analyze the opinion of the participants on some questions which could be of specific interest for the university.

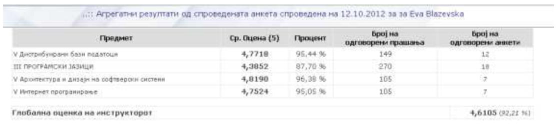


.:: Агрегатни резултати од спроведената анкета спроведена на 10.12.2012

Прашање	Ср. Оцена	Број на анкетни ливчиња
1. Case studies and assignments were applicable and useful	4,3459	532
2. How would you rate the content of the course?	4,4251	534
3. Lectures involve working in groups and student-group interaction	4,2444	532
4. Students were actively involved in the learning-teaching process	4,3352	533
5. The content of the course matched the course description	4,5768	533
6. The course contained the appropriate challenging college level	4,3981	530
7. The course was well organized and delivered full content	4,3977	533
8. The group size during the course of the semester was adequate	4,5628	532
9. The instructor always arrives and starts the lectures on time	4,7453	534
10. The instructor treats and respects individual rights equally	4,6642	533
11. The instructor uses the standard University grading policy	4,7148	533
12. The instructor was well prepared for each class	4,6679	533
13. The syllabus clearly described course assignments, dates & expectations	4,6635	532
14. The syllabus was delivered during the first day of class	4,7121	535
15. What is the quality of the handout materials/books/ pp-presentations?	4,3722	532
16. Would you recommend or take another course from the same instructor?	4,3959	533
17. Големината на групата беше соодветна	4,5284	2.377
18. Дали би земале уште еден предмет од истиот наставник	4,2938	2.379
19. Како би ја оцениле содржината на предметот?	4,3347	2.381
20. Како е квалитетот на поделените материјали за учење?	4,1198	2.367
21. Курсевите и домашните беа корисни за совладување на материјата	4,2651	2.369
22. Наставата вклучуваше групна настава и интерактивност	4,2541	2.381
23. Наставникот ги почитува индивидуалните права на студентите	4,5495	2.382
24. Наставникот користи стандардна УАКС шема за оценување	4,5628	2.379
25. Наставникот редовно присуствува и почнува со настава на време	4,5937	2.380
26. Предметот беше добро организиран и целосно предаден	4,2891	2.383
27. Предметот беше со соодветна академска тежина	4,3281	2.374
28. Професорот беше добро подготвен за предметот и часовите	4,4832	2.384
29. Силабусот беше поделен на првиот час од наставата	4,6213	2.382
30. Силабусот јасно прикажува содржината и очекувањата од предметот	4,4898	2.376
31. Содржината на предметот е идентична со опис на предметот	4,4342	2.376
32. Студентите беа активно вклучени во образовниот процес	4,2613	2.384

Fig. 8. Total user satisfaction per survey

Aggregate report for instructors' work. Although quite simple, this is probably the most important report in the process of evaluation of the academic staff. This report gives a clear view of the overall student satisfaction for that instructor in a defined period. If we combine this report with the detailed teacher-course reports we can create a powerful tool for detailed assessment of the academic staff.



.:: Агрегатни резултати од спроведената анкета спроведена на 12.10.2012 за за Eva Blazevska

Предмет	Ср. Оцена (5)	Процент	Број на одговорени прашања	Број на одговорени анкети
V Дистрибуирани бази податоци	4,7718	95,44 %	149	12
III ПРОГРАМСКИ ЈАЗИЦИ	4,3852	87,70 %	270	18
V Архитектура и дизајн на софтверски системи	4,8190	96,38 %	105	7
V Интернет програмирање	4,7524	95,05 %	105	7

Глобална оценка на инструкторот	4,6105 (92,21 %)

Fig. 9. Agregate report for a teacher

Aggregate report for school/university. This report is also used for measuring the level of satisfaction of the students but this time the results are grouped based on the department they describe. The questions that are part of the questionnaire are categorized but the differentiation is not visible. When the results are processed, a predefined methodology is used to group the grades by different categories. Each category then represents different department. By analyzing this report the coordinator could have a clear view about the quality of the work of different departments on a faculty or university level

D. Storage and further use of the accumulated data

If reduction of cost and time consumption are defined as the two largest benefits that the system for electronic evaluation can provide then probably the third largest benefit of this system would be the method of data storage. While the traditional method of evaluation stores huge amount of hard copy documents, this system uses modern database management system for storage and reuse of the collected data. Each result from the evaluation process is stored in the database and can be used whenever needed. This provides for creation of cross reports for multiple time periods which will enable the coordinator to monitor the evaluation process in general (Fig. 9&10) and follow the quality of teaching on a particular level (Fig. 11).

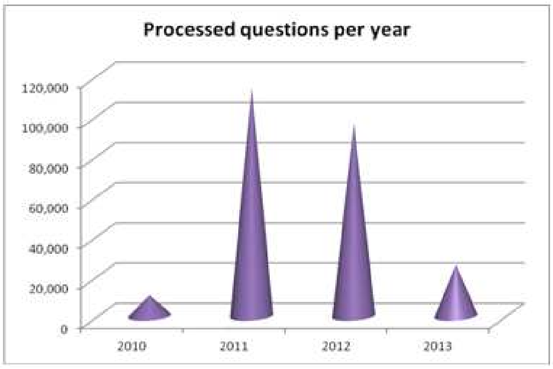


Fig. 10. Processed questions per year

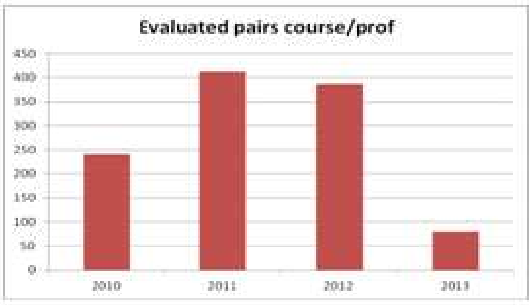


Fig. 11.a. Evaluated pairs course/professor

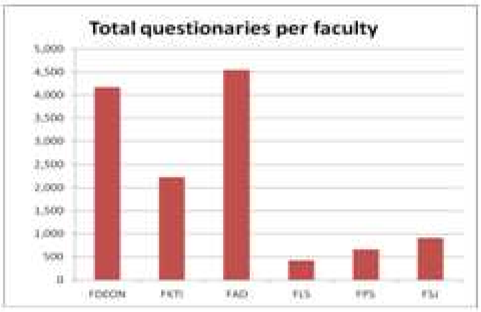


Fig. 12.b. Total questionnaires per faculty

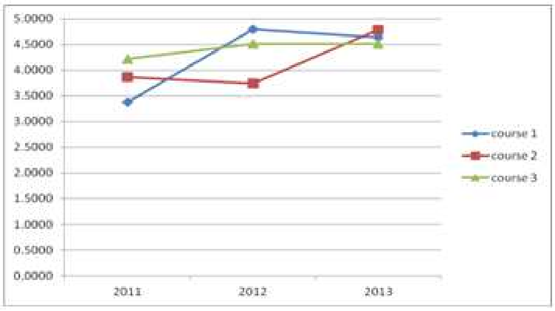


Fig. 13. Three courses (by three teachers) – average grades per course per generation

E. *Security aspect*

The implementation of the model is based on the web architecture. Since it has dedicated audience, it does not have to have public IP address, but can be a part of the institution intranet.

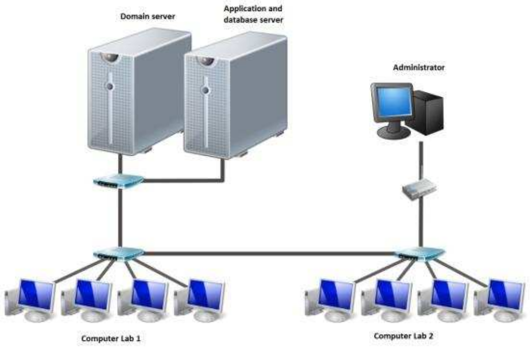


Fig. 14. System’s architecture

This is the first level of security, since the system can not be used outside the institution. Further on, in order to achieve even better security for the data, the system has several security policies for accessing its modules. Every system access is monitored and recorded into log file that is processed in special system back-end module. Specific modules of the system can be accessed only with special authorization mode and only from specific machines from the local network of the institution.

When filling the surveys, the system ensures that every questionary, although anonimusly, can be filled only once by a user. Complex system of cookies, sessions and active directory parameters are used so that the procees is anonimus, but ensures that the user can not manipulate or misuse the system.

III. CONCLUSION AND FUTURE WORK

So far, there were 1.200 polls with 353.000 questionnaires filled. This means that the same number of sheets was saved, not taking into account the time saved for data processing.

The implemented model proved to be flexible and easily adapchart for various surveys. One click and the polls for all the professors and courses are generated. Also, the questionnaires are clear and precise; the reports are comprehensive; there is no possibility for the breach of security.

The historical aspect of the data allows for implementation of data mining techniques, which can result in unexpected discoveries, which will enable further improvement of the evaluation process.

Finally, considering the more modern aspects of 360 degree evaluation and its continual evolution, appropriate surveys can be created to achieve desired goals.

REFERENCES

[1.] Cheung, G. (1999) Introducing a 360 degrees performance evaluation. Strategic change , 8 (2), p111-117

[2.] Fleenor, J. W., & Prince, J. M. (1997).Using 360-degree feedback in organizations: An annotated bibliography. Greensboro, NC: Center for Creative Leadership

[3.] Edwards, M. R., & Ewen A. (1996). 360° Feedback: the Powerful New Model for Employee Assessment & Performance Improvement. New York: AMACOM.

[4.] Ortega S., Baptiste L. & Beauchemin A. (2008) A Model for 360 Degree Teacher Evaluation in the Context of the CSME. Accessed online at: https://www.academia.edu/577036/A_model_for_360_degree_teacher_evaluation_in_the_context_of_the_CSME (last accessed on April 2, 2014)

HTML5 based Facet Browser for SPARQL Endpoints

Martina Janevska, Milos Jovanovik, Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia

Abstract—**The Linked Data concept uses a collection of Semantic Web technologies in order to interconnect, publish, and share pieces of data on the Web, in a machine-readable format. It enables querying and combining data from different datasets together in order to retrieve specific information and enable use-case scenarios which are unavailable over isolated datasets. However, the process of querying linked data published on different places on the Web poses several challenges. Generally, each user should know the schema of the data, write a query and access a relevant linked data endpoint. According to statistics, endpoints are often unavailable and have significant downtime, so this presents a serious obstacle in application development and scenarios which rely on the data. In this paper, we present a facet browser for SPARQL endpoints, based on HTML5. It allows users to search and retrieve RDF triples based on a keyword, from public SPARQL endpoints. By using HTML5 Web Storage, the triples from the results can be saved in the browser, locally, for future use. The facet browser provides management functionalities over the stored data - capabilities to update, refresh, modify, delete and download the triples in various RDF formats: JSON-LD, Turtle, NTriples, RDF/XML, JSON, CSV. The locally stored RDF triples can also be shared with other users. We believe that these features of the facet browser will help overcome the endpoint downtime issues, by providing offline data accessibility for the user and his applications.**

Keywords—Facet Browser; Linked Data; SPARQL; HTML5; Semantic Web;

I. INTRODUCTION

The Linked Data principles propose using the Web to create typed links between data from different sources. These may be as diverse as databases maintained by two organizations in different geographical locations, or simply heterogeneous systems within one organization that, historically, have not easily interoperated at the data level [1]. W3C provides a palette of technologies (RDF, GRDDL, POWDER, RDFa, the upcoming R2RML, RIF, SPARQL) to get access to the data[1]. SPARQL is a query language that is flexible and is mostly used for interacting with RDF databases also known as triple stores. As a standardized query language it contains a lot of specifications, query operations, request and response formats.

Every dataset declared as Linked Data can be accessed using the SPARQL language and queries distributed over different SPARQL endpoints. For retrieving the data, a user has to access available endpoints and to be able to query them. The ability to query the data space of Linked Data provides benefits which have not been possible before. Data from different data sources can be gathered, and scattered information from multiple sources can be joined in order to achieve a more complete view of a given domain and support more complex use-case scenarios.

Executing queries over the Web of Linked Data poses several challenges that do not arise in traditional query processing. Due to the openness of the data space, it is not possible to know all data sources that might be relevant for answering a query in advance. People who need some data from the Linked Data cloud[2] sometimes do not know how the specified data is represented in datasets and they face the challenge to write complex queries in order to extract results.

Another issue which arises in the domain of querying data over the data space of Linked Data is the availability of the SPARQL endpoints. According to statistics[3], these endpoints are often unavailable and have significant downtime. This presents a serious obstacle in developing ‘killer applications’ over Linked Data datasets, and forces developers to focus their time and energy on other technologies and approaches.

In this paper, we present a facet browser for SPARQL endpoints, which tries to overcome these two main issues. Firstly, it offers keyword lookup in RDF datasets from SPARQL endpoints, mitigating the issue of schema knowledge beforehand, or knowing the SPARQL language at all. Secondly, it allows storage of the results received from a SPARQL endpoint, for future use. The results are being stored in HTML5 Web Storage, in the user browser, and can be manipulated by the user. The data can also be serialized into all common RDF formats, and shared with others.

II. RELATED WORK

Browsers can be a crucial tool when someone is dealing with data from the Linked Data cloud. They should provide different use-cases and options for users and their needs. That is the main reason why there are a number of browser applications for Linked Data and they all provide different approaches in dealing with data and presenting results to the end user.

[1] http://www.w3.org/standards/semanticweb/data

[2] http://lod-cloud.net/

[3] http://sparqles.okfn.org/availability

The Tabulator project[4] represents a generic data browser, which provides ways for browsing RDF resources published on the Web, and follow RDF links from one resource to another [2]. The main goal of Tabulator is to increase the usage of Linked Data, explore the potentials and restrictions of the Semantic Web architecture and to increase development in the field of generic data interfaces.

The Disco - Hyperdata Browser[5] is a browser which is used for handling the Semantic Web as an unbounded set of data sources. This tool renders all information related to a resource, which is specified by his URI entered into the navigation box. For this resource, the user gets a description which contains hyperlinks or facets allowing him to navigate between resources. When a user moves from one resource to other, the browser dynamically retrieves and displays information as a property-value table. Also, the browser stores all retrieved RDF graphs, whose hyperlinks have been clicked, in a session cache and provide an option for showing them in a list in a new browser window.

Swoogle[6] is a specialized web based data browser used for discovering, analyzing and indexing of data from datasets published on the Web with Semantic Web technologies. Swoogle explains and records significant metadata about data published in these datasets and their fundamental parts (e.g., terms, individuals, triples). As a browser it uses its search and navigation services in order to provide scalable service for accessing Semantic Web data and finding relevant documents [3]. On one hand, similar to our solution, Swoogle represents a keyword-based search engine and it does not require schema knowledge and query language expertise. That is an important reason why it is really appropriate for non-technical users. On the other hand, similarly as Tabulator and Disco – Hyperdata Browser, Swoogle does not provide any options for further management and wider usage of retrieved RDF results, which is one of main focuses used in our approach.

Longwell[7] is a web-based faceted browser, considered as a combination of the flexibility of the RDF data model and the effectiveness of the faceted browsing paradigm. It is a powerful tool, and just like our solution it enables visualization and browsing complex RDF datasets, allows the user to quickly get an overview of what data is present in the dataset, and offers specific information about resources.

Virtuoso Faceted Browser[8] is keyword-based faceted browser and as a solution it is the most advanced. The user can enter a text search pattern and get a results page containing a list of literal value snippets from property values associated with the searched pattern. With a click on some of the entities or relations, the user gets new results with description of that particular object. There is also an option for getting raw data from search result in several formats, such as CSV, JSON, XML, N-Triples, etc.

The facet browser we have developed has several additional options which differentiate it from all of these examples. The user of our applications gets the results for the used keyword in a table, and has the additional option to save them for later use. The saved data can be managed by the user; it can be updated and refreshed, modified, removed, and serialized on the local machine in various RDF formats. The local storage of RDF data, in the browser, allows other future features, as well.

III. THE FACET BROWSER

The facet browser for SPARQL endpoints (Fig. 1) is a HTML5 web-based application which has a main focus of searching and retrieving RDF triples which contain a given keyword, from a specified SPARQL endpoint. The results are presented in a human-readable representation. The endpoint does not require any SPARQL or RDF knowledge from the user.

A. *Data Retrieval*

One of the main functionalities in our application is the effective search for data from a SPARQL endpoint. There are two basic things which should be provided by the user to the application: a search term and SPARQL endpoint URL. The application takes the search term of the user as a keyword, and sends a SPARQL query to the endpoint provided by the user.

In order to get these results, we use the following SPARQL query, in which we replace “pattern” with the keyword of the user:

```
SELECT ?s ?p ?o
WHERE
{
      ?s ?p ?o.
      FILTER ( regex(?o, "pattern", "i"))
}
LIMIT 1000
```

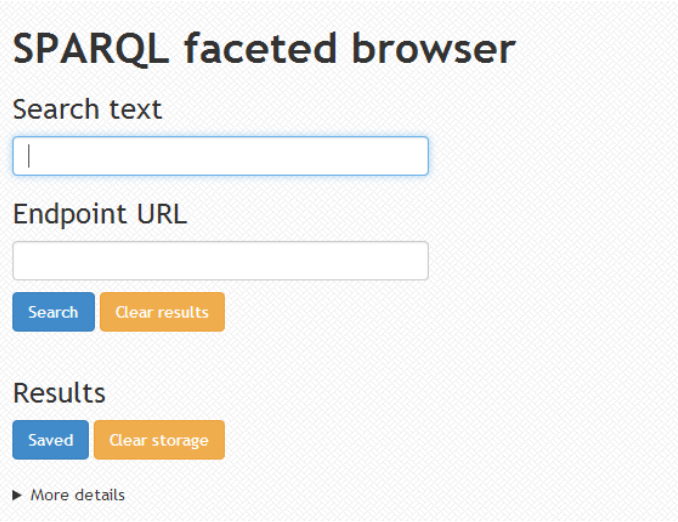


Fig. 1. The facet browser interface.

[4] http://www.w3.org/2005/ajar/tab
[5] http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/disco/
[6] http://swoogle.umbc.edu/
[7] http://simile.mit.edu/wiki/Longwell
[8] http://lod.openlinksw.com

With the use of this specific query, we get a number of triples, composed of literal value snippets from property values related with the searched keyword.

In the query we place a limit of maximum 1000 triples, in order to comply with the fair use policy of Linked Data. The results are shown in a table with three columns: entity, relation and value, sorted by entity values, and with the possibility of displaying a particular number (10, 50, 100 or 1000) of result rows per page (Fig. 2).

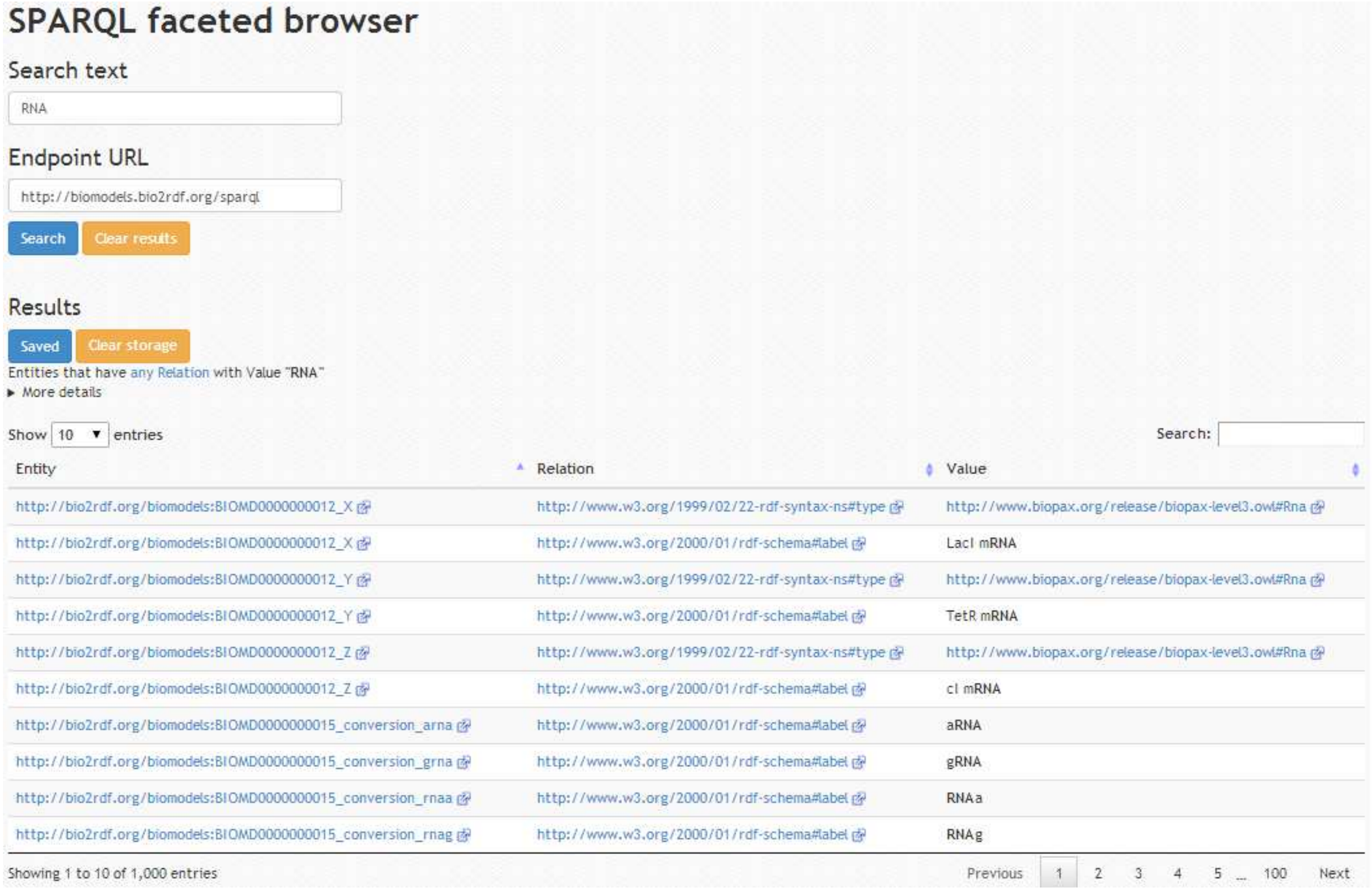
Faceted navigation is an examining technique which uses several different ways for navigating through a collection of components, instead of using a specific and determined order. Facet browser interfaces provide navigation through collections of data in a user-friendly way [4]. When users are facing Linked Data datasets, they often are not sure what exactly they are looking for, so they may not always be familiar with the domain and its schema. Or, the users may just want to learn something new about a particular matter.

With our faceted browser we allow users to search data through multiple successive iterations without having excessive (or any) knowledge of the datasets. In the results table, the entities and the relations are displayed as facets, providing a multi resource scheme. The users can click on the any of the facets, and get a description of the resource or relation. This is done by the application by sending a new query to the endpoint, asking for the description of that resource or relation. As a result from the 'describe' query, the browser displays a table with two columns: relation and value, displaying all of the relations and object values related to the entity (Fig. 3).

There are also options for the users to see more details about the displayed result data, and get all of the results from the endpoint, not just the first 1000. This supports use-cases for more advanced users, who can be interested in more technical details about the query, or would like to get all possible RDF triples for their initial query.

The application also has the capability of saving data. We use HTML5 for the web application, which provides us with the Web Storage functionality. This functionality supports persistent data storage with great capacity. It offers two different types of storage: local and session. For our application we use the former, because it stores data with no expiration date, which means the data will not be deleted and will persist when the user closes the browser and his session ends. The application allows the users to define details about the RDF triples before saving them in Web Storage (Fig. 4).

This Web Storage functionality allows our application to provide offline access to the RDF results from the saved queries. This is the feature which helps Linked Data users mitigate the issue of having the needed SPARQL endpoint available at any given moment. This client-side storage is a programming model which does not require a server infrastructure and single user can only access his local files, which means that our web application becomes personalized.



SPARQL faceted browser

Search text

RNA

Endpoint URL

http://biomodels.bio2rdf.org/sparql

Results

Entities that have any Relation with Value "RNA"

▸ More details

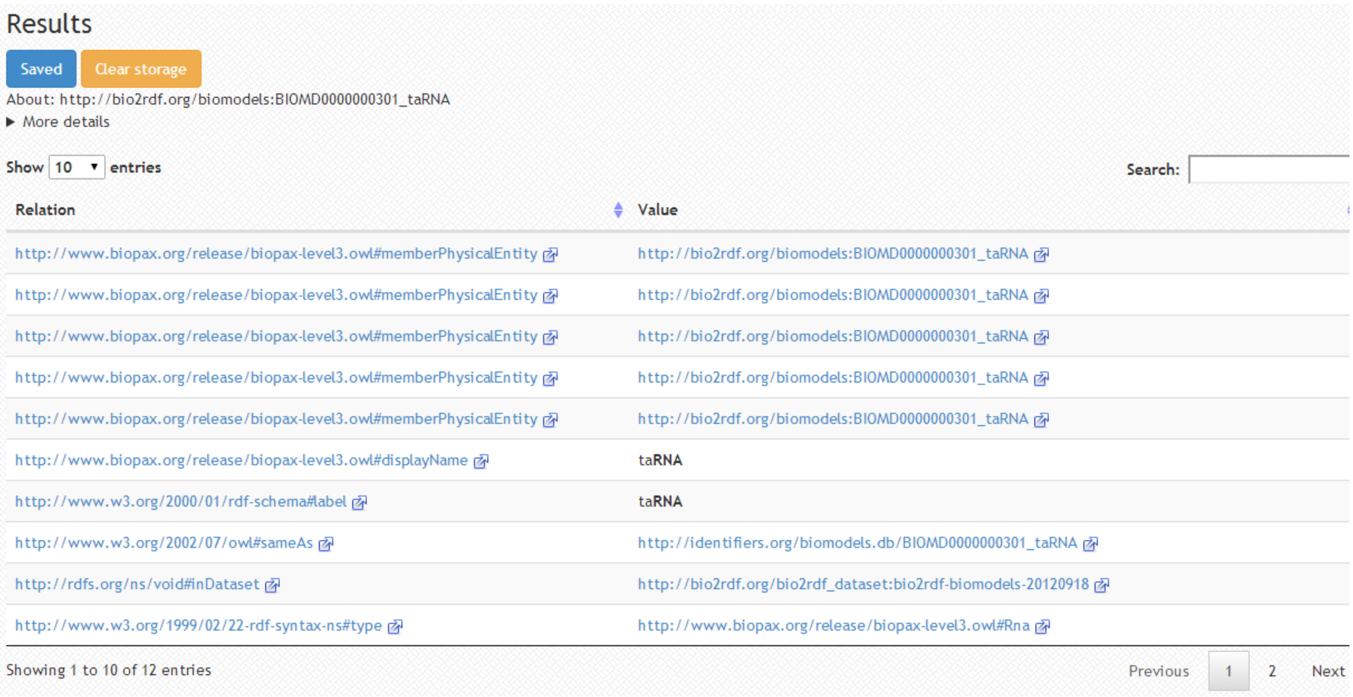
Show 10 ▾ entries

Search:

Entity	Relation	Value
http://bio2rdf.org/biomodels:BIOMD0000000012_X	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.biopax.org/release/biopax-level3.owl#Rna
http://bio2rdf.org/biomodels:BIOMD0000000012_X	http://www.w3.org/2000/01/rdf-schema#label	LacI mRNA
http://bio2rdf.org/biomodels:BIOMD0000000012_Y	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.biopax.org/release/biopax-level3.owl#Rna
http://bio2rdf.org/biomodels:BIOMD0000000012_Y	http://www.w3.org/2000/01/rdf-schema#label	TetR mRNA
http://bio2rdf.org/biomodels:BIOMD0000000012_Z	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.biopax.org/release/biopax-level3.owl#Rna
http://bio2rdf.org/biomodels:BIOMD0000000012_Z	http://www.w3.org/2000/01/rdf-schema#label	cI mRNA
http://bio2rdf.org/biomodels:BIOMD0000000015_conversion_arna	http://www.w3.org/2000/01/rdf-schema#label	aRNA
http://bio2rdf.org/biomodels:BIOMD0000000015_conversion_grna	http://www.w3.org/2000/01/rdf-schema#label	gRNA
http://bio2rdf.org/biomodels:BIOMD0000000015_conversion_rnaa	http://www.w3.org/2000/01/rdf-schema#label	RNAa
http://bio2rdf.org/biomodels:BIOMD0000000015_conversion_rnag	http://www.w3.org/2000/01/rdf-schema#label	RNAg

Showing 1 to 10 of 1,000 entries

Fig. 2. Results from searching the keyword 'Aspirin'.



Results

About: http://bio2rdf.org/biomodels:BIOMD0000000301_taRNA

▶ More details

Show 10 ▾ entries

Search:

Relation	Value
http://www.biopax.org/release/biopax-level3.owl#memberPhysicalEntity	http://bio2rdf.org/biomodels:BIOMD0000000301_taRNA
http://www.biopax.org/release/biopax-level3.owl#memberPhysicalEntity	http://bio2rdf.org/biomodels:BIOMD0000000301_taRNA
http://www.biopax.org/release/biopax-level3.owl#memberPhysicalEntity	http://bio2rdf.org/biomodels:BIOMD0000000301_taRNA
http://www.biopax.org/release/biopax-level3.owl#memberPhysicalEntity	http://bio2rdf.org/biomodels:BIOMD0000000301_taRNA
http://www.biopax.org/release/biopax-level3.owl#memberPhysicalEntity	http://bio2rdf.org/biomodels:BIOMD0000000301_taRNA
http://www.biopax.org/release/biopax-level3.owl#displayName	taRNA
http://www.w3.org/2000/01/rdf-schema#label	taRNA
http://www.w3.org/2002/07/owl#sameAs	http://identifiers.org/biomodels.db/BIOMD0000000301_taRNA
http://rdfs.org/ns/void#inDataset	http://bio2rdf.org/bio2rdf_dataset:bio2rdf-biomodels-20120918
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.biopax.org/release/biopax-level3.owl#Rna

Showing 1 to 10 of 12 entries

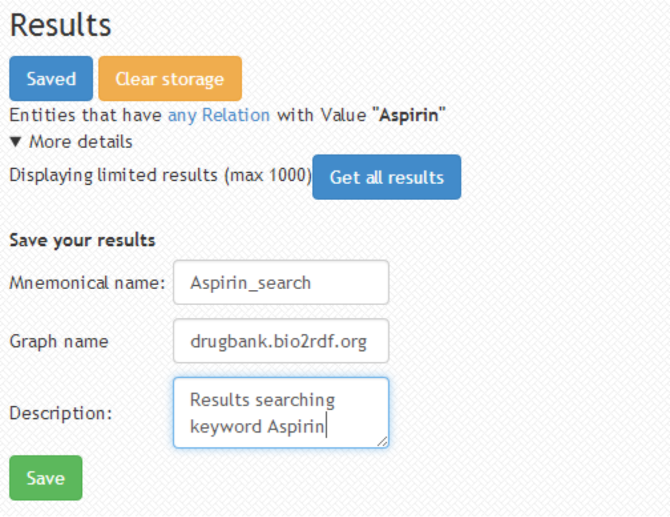
Fig. 3. A resource description, showing its relations and values.

After a result set is saved, the user can then manage it, or continue querying a SPARQL endpoint.

B. Storage Management

As we already mentioned, the application allows the users to access and manage their stored data. An example stored dataset is shown on Fig. 6. Each saved dataset, aside from the RDF triples, contains a mnemonic name, graph name and a description, which are used to distinguish the datasets.

The application provides management options to the users for manipulating the RDF triples from a stored dataset (Fig. 6). The 'Details' option allows a user to see all of triples which are parts of that dataset, instead of only the first 10. A user can also edit the mnemonic name, the graph name and the description (Fig. 5).



Results

Entities that have any Relation with Value "Aspirin"

▼ More details

Displaying limited results (max 1000)

Save your results

Mnemonical name: Aspirin_search

Graph name drugbank.bio2rdf.org

Description: Results searching keyword Aspirin

Fig. 4. Form for saving data in local storage.

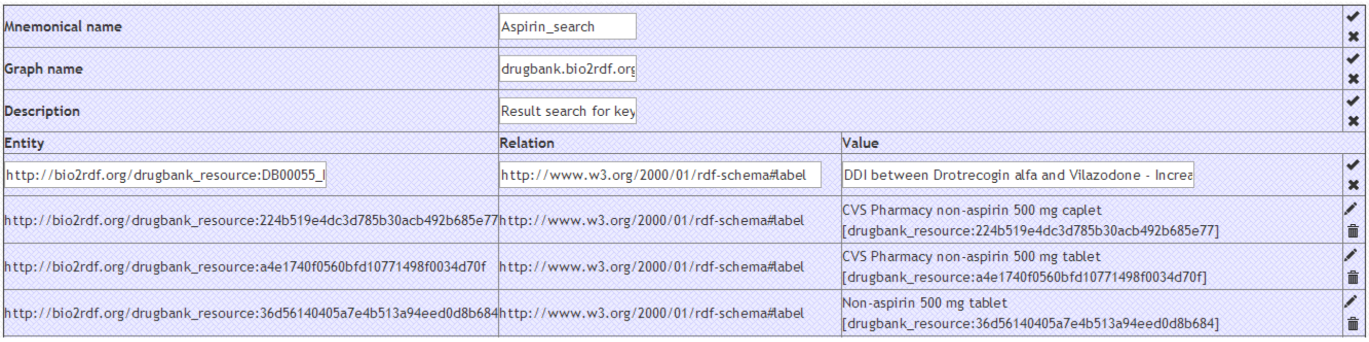
However, one of the main functionalities over the stored datasets is the ability to edit the RDF triples themselves (Fig. 5). This is also a feature available in the application, and its intention is to provide the user with the ability to modify the datasets in order to correct, change or delete some of the RDF triples which he intends to use for other purposes.

In order to provide the capabilities of storing the dataset triples as into HTML5 Web Storage, as well as review and manage them, we use the JavaScript wrapper library triplestore.js[9]. The library offers several functionalities which provide RDF data administration for the user. Additionally, it is a library which is very simple to use and develop with.

When a user has an RDF dataset stored in his browser, it can only be viewed and manipulated through the application. However, if the user needs a dataset for other purposes, such as using it as a data layer for an application, or using it in data analysis, the user may need to serialize the data and save it locally on his machine.

Our application offers this functionality to the users, by allowing serialization of a saved dataset in various RDF formats: JSON-LD, Turtle, NTriples, RDF/XML and CSV. The serialized data is showed in the web application, for supporting use-cases when the user would like to copy and paste the dataset content, but the application also allows for the serialized RDF dataset to be downloaded as a file, on the local machine (Fig. 7).

[9] http://www.w3.org/2013/04/semweb-html5/triplestoreJS/



Mnemonical name	Aspirin_search		✓ ✖
Graph name	drugbank.bio2rdf.org		✓ ✖
Description	Result search for key		✓ ✖
Entity	**Relation**	**Value**	
http://bio2rdf.org/drugbank_resource:DB00055_I	http://www.w3.org/2000/01/rdf-schema#label	DDI between Drotrecogin alfa and Vilazodone - Increa	✓ ✖
http://bio2rdf.org/drugbank_resource:224b519e4dc3d785b30acb492b685e77	http://www.w3.org/2000/01/rdf-schema#label	CVS Pharmacy non-aspirin 500 mg caplet [drugbank_resource:224b519e4dc3d785b30acb492b685e77]	✎ 🗑
http://bio2rdf.org/drugbank_resource:a4e1740f0560bfd10771498f0034d70f	http://www.w3.org/2000/01/rdf-schema#label	CVS Pharmacy non-aspirin 500 mg tablet [drugbank_resource:a4e1740f0560bfd10771498f0034d70f]	✎ 🗑
http://bio2rdf.org/drugbank_resource:36d56140405a7e4b513a94eed0d8b684	http://www.w3.org/2000/01/rdf-schema#label	Non-aspirin 500 mg tablet [drugbank_resource:36d56140405a7e4b513a94eed0d8b684]	✎ 🗑

Fig. 5. Local storage data management.

Because of the dynamic nature of Linked and Open Data datasets, a user may want to update a locally saved resulting RDF dataset, once the source SPARQL endpoint is available.

In order to provide support for this scenario, we added the functionality of refreshing the data from a stored RDF dataset. In order to do this, we store the initial query used for obtaining the dataset, along with the source SPARQL endpoint, so when a user needs to refresh the data, we send a new query to the SPARQL endpoint and retrieve the updated results. After a dataset refresh, we prompt the user in order to see whether he wants to do an update refresh or a clean refresh. If he chooses the former, only the newly retrieved triples which did not exist in the previously stored results are added to the RDF dataset, and if he chooses the latter, the newly retrieved results will replace the previously stored ones.

Since HTML5 Web Storage has different storage capacity for local storage depending on the user web browser and RDF results retrieved from an endpoint can reach the size of thousands of triples, we also provide the option for deleting an RDF dataset from the web browser storage (Fig. 6).

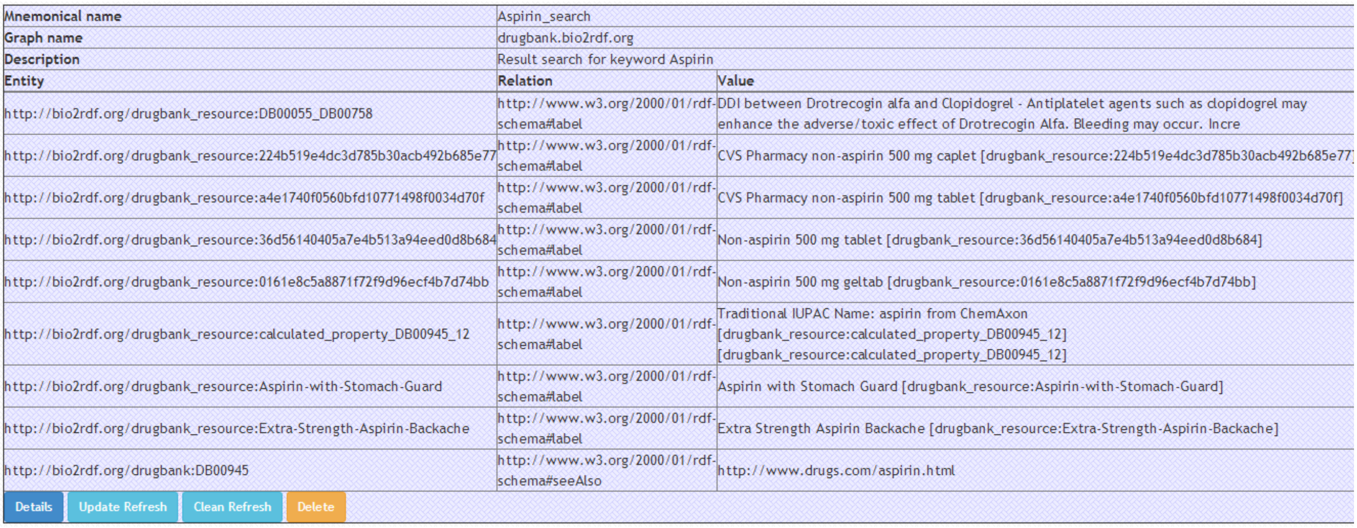
C. *About the application*

The Web of Linked Data is open, so applications which work with Linked Data can follow relations between data and learn about new data sources. As a result of that, with increase of published data, these applications will provide to the users more complete information and knowledge [5]. This also means that anyone can publish data structured with these technologies on the Web, and become part of the increasingly growing LOD cloud. With this, every Linked Data publisher adds more possible use-cases over the interconnected datasets on the Web, and with this, indirectly adds more value to the LOD cloud datasets.

Therefore, we based our HTML5 facet browser on the same principles of openness and collaborative value leverage. We developed and published it as an open-source application[10] where other can also participate in further development, and deployed a public instance[11].

IV. CONCLUSION AND FUTURE WORK

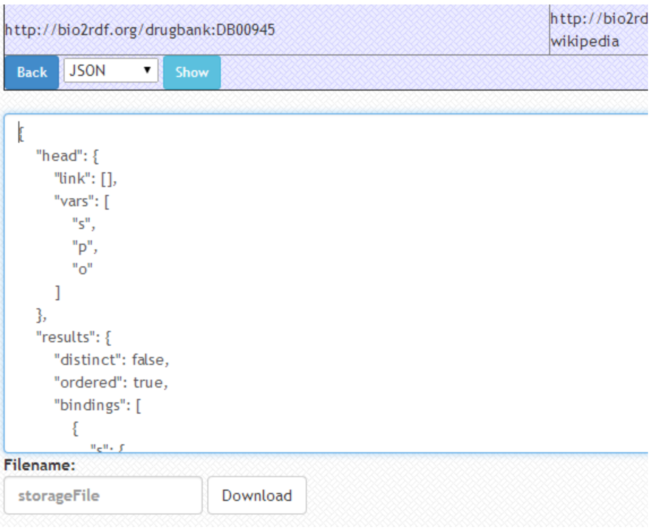
In general, Semantic Web technologies are used for publishing and managing Open Data, and interconnecting it with other data in the Linked Data cloud; using already



Mnemonical name	Aspirin_search	
Graph name	drugbank.bio2rdf.org	
Description	Result search for keyword Aspirin	
Entity	**Relation**	**Value**
http://bio2rdf.org/drugbank_resource:DB00055_DB00758	http://www.w3.org/2000/01/rdf-schema#label	DDI between Drotrecogin alfa and Clopidogrel - Antiplatelet agents such as clopidogrel may enhance the adverse/toxic effect of Drotrecogin Alfa. Bleeding may occur. Incre
http://bio2rdf.org/drugbank_resource:224b519e4dc3d785b30acb492b685e77	http://www.w3.org/2000/01/rdf-schema#label	CVS Pharmacy non-aspirin 500 mg caplet [drugbank_resource:224b519e4dc3d785b30acb492b685e77]
http://bio2rdf.org/drugbank_resource:a4e1740f0560bfd10771498f0034d70f	http://www.w3.org/2000/01/rdf-schema#label	CVS Pharmacy non-aspirin 500 mg tablet [drugbank_resource:a4e1740f0560bfd10771498f0034d70f]
http://bio2rdf.org/drugbank_resource:36d56140405a7e4b513a94eed0d8b684	http://www.w3.org/2000/01/rdf-schema#label	Non-aspirin 500 mg tablet [drugbank_resource:36d56140405a7e4b513a94eed0d8b684]
http://bio2rdf.org/drugbank_resource:0161e8c5a8871f72f9d96ecf4b7d74bb	http://www.w3.org/2000/01/rdf-schema#label	Non-aspirin 500 mg geltab [drugbank_resource:0161e8c5a8871f72f9d96ecf4b7d74bb]
http://bio2rdf.org/drugbank_resource:calculated_property_DB00945_12	http://www.w3.org/2000/01/rdf-schema#label	Traditional IUPAC Name: aspirin from ChemAxon [drugbank_resource:calculated_property_DB00945_12] [drugbank_resource:calculated_property_DB00945_12]
http://bio2rdf.org/drugbank_resource:Aspirin-with-Stomach-Guard	http://www.w3.org/2000/01/rdf-schema#label	Aspirin with Stomach Guard [drugbank_resource:Aspirin-with-Stomach-Guard]
http://bio2rdf.org/drugbank_resource:Extra-Strength-Aspirin-Backache	http://www.w3.org/2000/01/rdf-schema#label	Extra Strength Aspirin Backache [drugbank_resource:Extra-Strength-Aspirin-Backache]
http://bio2rdf.org/drugbank:DB00945	http://www.w3.org/2000/01/rdf-schema#seeAlso	http://www.drugs.com/aspirin.html

Fig. 6. Table of data saved in local storage.

[10] https://github.com/mjanevska/html5-sparql-browser
[11] http://fct.linkeddata.finki.ukim.mk



```
{
  "head": {
    "link": [],
    "vars": [
      "s",
      "p",
      "o"
    ]
  },
  "results": {
    "distinct": false,
    "ordered": true,
    "bindings": [
      {
        "s": {
```

Filename: storageFile

Fig. 7. Data serialization and download.

published Open and Linked Data from public SPARQL endpoints; developing applications based on Open and Linked data, either stored locally, or available on the Web.

Our HTML5 facet browser is intended to be used as a search engine for datasets published as Linked Data, available through SPARQL endpoints. It provides a flat browsing capability through the dataset by combining search, facets and operations on sets of resources.

Using HTML5 Web Storage, we enable the option to store the RDF results for future use, locally, in the browser storage. The facet browser provides a broader view on the stored data and management functionalities over it. The locally stored RDF triples can be serialized in several RDF formats, downloaded as separated files and then used in applications, analysis or shared with other users. We believe these features help the users to overcome the existing SPARQL endpoint availability issues.

The original goal of the project was to simplify the process of searching the Web of Open and Linked Data by using facets, and to provide more services and options to their users. With our application, we would also like to encourage development of innovative applications and services over Linked and Open Data.

The future development of the project will include a feature for providing more transparent sharing of the stored RDF datasets. We plan to add an option for publishing a dataset in a chosen RDF format on a permanent URL, in order to support use-cases in which a user would like to share an RDF dataset with other, via the Web.

We also plan to add more management options for the stored data. By using other functionalities from triplestore.js, we can enable scenarios in which the users would filter a stored RDF dataset and obtain smaller subsets of it. These filtered subsets will contain RDF triples with subjects which are matched with an optional property and value, or triples which contain values that match with a specified subject and optional property.

ACKNOWLEDGEMENT

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje, as part of the research project “Semantic Sky 2.0: Enterprise Knowledge Management”.

REFERENCES

[1] Bizer, C., Heath, T., Berners-Lee, T.: “Linked Data - the story so far”. Journal on Semantic Web and Information Systems, 2009.

[2] Berners-Lee, T.: “Tabulator: Exploring and analyzing linked data on the semantic web”. Proceedings of the 3rd International Semantic Web User Interaction Workshop, 2006.

[3] Finin, T.: "Swoogle: Searching for knowledge on the Semantic Web" Proceedings, AAAI 05, 2005.

[4] Brunetti J. M., Gil R., Garcia R.: “Facets and Pivoting for Flexible and Usable Linked Data Exploration”. Interacting with Linked Data, 2012.

[5] Bizer, Christian. "The emerging Web of Linkedata." Intelligent Systems, IEEE 24.5, 2009.

ICT tools for Policy Design in City of Skopje

Marjan Gusev
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16, Skopje, Macedonia
Email: marjan.gushev@finki.ukim.mk

Goran Velkoski
Innovation Dooel,
Vostanichka 118, Skopje, Macedonia
Email: goran.velkoski@innovation.com.mk

Aleksandar Avukatov, Mario Ringov, Lovren Markic, Mirjana Apostolova
City of Skopje,
Ilinden 82, Skopje, Macedonia
Email: {aleksandar.avukatov, mario.ringov, lovren.markic, mirjana.apostolova}@skopje.gov.mk

***Abstract*—This paper presents several outputs of FUPOL FP7 project aiming at implementing sophisticated ICT tools to support policy modeling processes for urban development of City of Skopje. The administration of City of Skopje is actively participating along with Innovation Dooel to enable usage of social networking tools, optimization, simulation and visualization tools in policy design and public involvement. The project goal is to develop FUPOL social networking policy model and ICT platform. It aims at developing several tools that optimize, simulate and visualize various scenarios within urban design and development of Vodno recreation facilities and introducing bike inter-modality. This paper discusses the tools implemented at City of Skopje and its benefits.**

***Index Terms*—FUPOL, FP7, Social networking, Optimization, Simulation, Visualization, City of Skopje.**

I. Introduction

In the past 2 years the City of Skopje and Innovation Dooel are participating in the project Intelligent Tools for Policy Design financed by the European Union, within the FP7 Program [1]. This project aims to provide a completely new approach to the traditional policies, and it's goal is to contribute towards improving the way politicians and administration communicate with the citizens [2].

The main output of the FUPOL project are software tools to be implemented for policy design and public involvement. Policy design tools can be supported in three parts:

- setting agenda for hot topics,
- analysis of policy design, and
- policy formulation.

Public involvement is fostered by activating several channels in the policy development process [3]. Participation is motivated by various social networking channels where campaigns are activated, and also by a set of social networking tools that measure and evaluate citizen responses. Social impact is obtained by conversational exchange, content amplification, sentiment, content appreciation, perceived influence etc.

A set of tools are also developed to support simulation, optimization and visualization [4]. These visual tools are offered to both the administration and the citizens. City of Skopje decided to concentrate on two projects: the first one considers Vodno recreational activities and the second fostering the bike inter-modality.

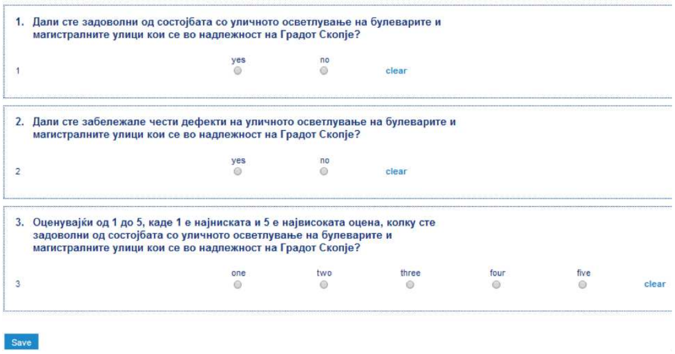
The rest of the paper is organized as follows: Section II presents the social networking tools implemented and in use at City of Skopje for policy design and modeling. In Section III we present the simulation and optimization recipes in use to develop particular projects, about Vodno recreational resources and fostering bike inter-modality, correspondingly in Section IV and Section V. We further discuss the benefits of use of these tools in Section VI, and in Section VII we give conclusions and discuss future implementation details.

II. Social Networking Tools

Basically, the FUPOL project output aims at developing FUPOL core platform, that is a software consisting of several software tools which will provide a direct line of communication between the citizens on one side, and the administration and politicians (decision makers) on the other, by using a variety of tools to monitor and collect data from the internet and social media. In case of City of Skopje the following channels are actively supported for communication with the citizens: Facebook, Twitter and Blogspot.

FUPOL core ICT platform supports social networking tools for FUPOL Policy Model, deemed appropriate to support the policy processes in the City of Skopje. This platform consists of the following ICT tools:

- *FUPOL Questionnaire*, a tool that realizes several questionnaire types essential to obtain citizen opinion. An example of use of this tool at City of Skopje is presented in Figure 1.
- *FUPOL Opinion map*, a tool that collects citizen opinion based on marking. labeling or commenting a certain pin in an interactive map. An example of use of an interactive opinion map is the one presented in Figure 2, where administration of City of Skopje got citizen responses where to implement recreation facilities at Vodno mountain.
- *FUPOL Social media window* is a tool that collects articles and posts in a social media aggregator. It has been tested with a number of sources in all languages including Macedonian with Cyrillic script (Twitter, Facebook, Blogs and RSS). The tool offers filtering and search features that enable the City of Skopje to analyze all the posts and news according to a certain selected topic. An



1. Дали сте задоволни од состојбата со уличното осветлување на булеварите и магистралните улици кои се во надлежност на Градот Скопје?

1 yes ○ no ○ clear

2. Дали сте забележале чести дефекти на уличното осветлување на булеварите и магистралните улици кои се во надлежност на Градот Скопје?

2 yes ○ no ○ clear

3. Оценувајќи од 1 до 5, каде 1 е најниската и 5 е највисоката оцена, колку сте задоволни од состојбата со уличното осветлување на булеварите и магистралните улици кои се во надлежност на Градот Скопје?

3 one ○ two ○ three ○ four ○ five ○ clear

Save

Fig. 1. An example of questionnaire used at City of Skopje blogspot

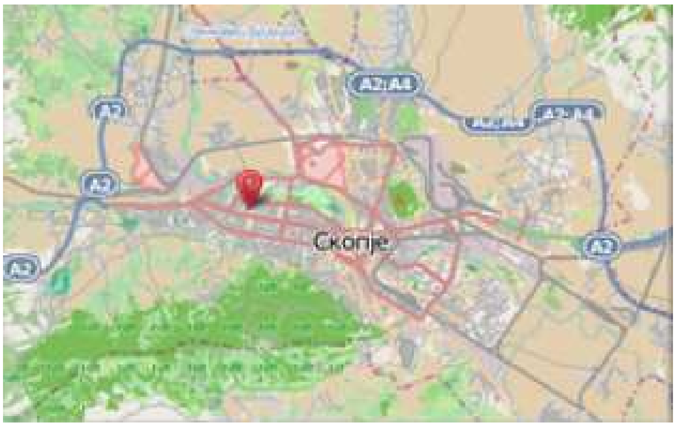


Fig. 2. An example of opinion used at City of Skopje Facebook

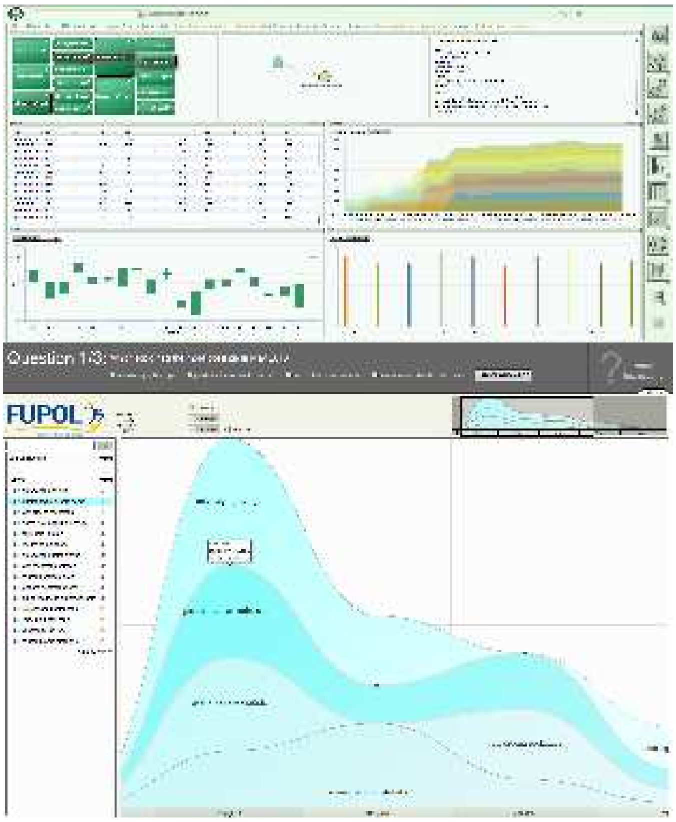


Fig. 3. An example of hot topic sensing features for trends

advanced search and analysis tool is under development by expanding the existing search criteria with enhanced Boolean searches, location based searches, and advanced filtering reducing irrelevant postings.

- *Heat Maps* is an advanced tool to identify locations of postings of a certain topic. It is a kind of a visualization tool that enables City of Skopje to find out locations where a certain issue has been discussed a lot or has been analyzed and commented. The function is considered very useful since it shows where specific opinions are coming from and helps to locate problem spots. However the number of postings with a given location is relatively small and to enable a good analysis more hits and postings should be present.
- *Hot Topic Sensing* tool is planned as an advanced tool for further development where a certain topic will be extracted as a hot topic discussed in forums and news [5].
- *Visualization tools* are those that support analyzes of previous news aggregating tools to collect and analyze topics from social networks and news aggregators. It adds a value to the hot topic sensing and heat maps by visualizing trends in topic discussions, and its application is presented in Figure 3. This tool is still under development and is expected in near future.

III. Simulation

The FUPOL project for City of Skopje consists of two separate projects. The first one targets the recreational area on Vodno Mountain and the second the Skopje bike tracks and parkings resources.

The main objective of the first project is the development of a software solution that will offer the City of Skopje and its citizens the opportunity to simulate the occupancy of the recreational resources on Vodno Mountain and therefore enable conclusions for better organization of recreational activities. The system would help the Administration of City of Skopje in improving the scheduling and resource planning, initiation and creating new projects involving the recreational area at Vodno Mountain. The citizens of City of Skopje will also be involved in the decision making process by continuos communication. They can rather easy express their opinion to the authorities, making the whole process more transparent and efficient.

Similarly the second project objective is the development of a software solution that will offer the opportunity to simulate the occupancy and usage of bike stations and bike parking lots. The overall goal is to increase the number of people that use bikes as a transport means, by taking several different measures, such as establishing bike inter-modality, initiating the development of parking lots, rent-a-bike facilities, new bike paths, improving existing bike paths.

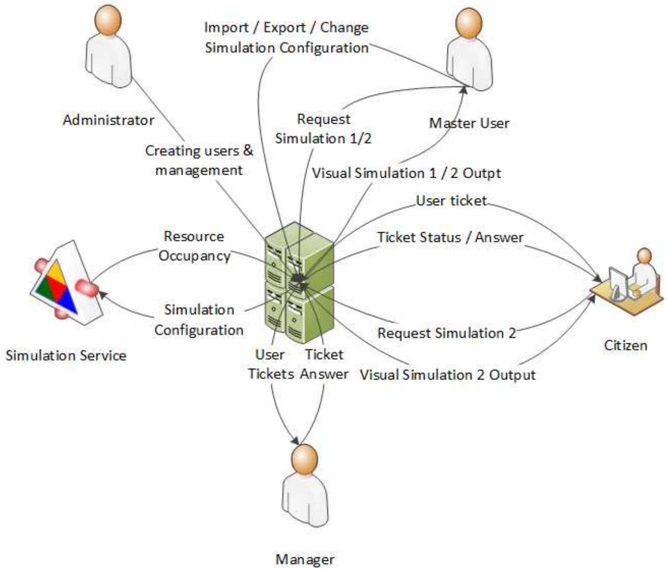


Fig. 4. Core simulation functionalities

The system would help the Administration of City of Skopje in improving the scheduling and resource planning, initiation and creation of new projects involving the bike stations and bike parking lots. The citizens of City of Skopje will also be involved in the decision making process by constantly communicating and expressing their opinion to the authorities, making the whole process more transparent and efficient.

A. *Functional Description*

The main functionality is the creation and execution of simulations. The core functions are presented in Figure 4 identified as:

- *simulation configuration definition* by importing, exporting and changing the existing sets of input information about resources and activities,
- *ticket management* realized by initiating tickets by citizens about new proposals for organization of recreation activities, and ticket responding by administration,
- *executing simulation* by running simulation and visualizing the results of resource occupancy.

More details about functionality will be explained in further subsections for description of users.

B. *Simulation types*

The purpose of the Simulation is to simulate resource occupancy and generate conflict reports for a given period of recreation activities on Vodno Mountain. Two simulations are proposed in the system:

- *Simulation 1* aiming at visualizing a one year period, intended for master users, where all weather conditions are simulated based on typical behavior
- *Simulation 2* with goal to visualize a one week behavior in a selected month with selected weather, intended for both master users and citizens

Simulation 1 is mainly to be performed by the administration of City of Skopje. The main idea is to find out which project can have the highest impact when implemented. In cases of limited budget, the City of Skopje has to decide which project to be implemented and therefore the use of simulation 1 is very important.

Simulation 2 is mainly offered to citizen to chick how will their proposals and re-organization ideas benefit in a simulated model. Users can reschedule certain events or change the number of users performing a certain activity and find out the resource occupancy.

C. *User roles*

The users in the simulation system are classified as the following user groups:

- *Administrator* creating and managing users.
- *Master User* setup (using import) of initial simulator configuration, change configuration, export configuration, and realize both simulation 1 and 2.
- *Manager* can list ticket reports, view and answer tickets sent by users, suggest new configuration and projects that will improve the recreation on Vodno Mountain or increase the number of users that bike as transport means.
- *Public user* can access the visual simulation interface for simulation 2.

D. *Basic visualization functionalities*

The Visualization part of the simulation software is intended to realize a visual representation of the generated data within the simulations. All data must be setup on a map of the region with clearly distinguishable resources and their locations.

There are visualization functionalities that are present on both projects. This functionalities are:

- **Map Navigation** functionality is available for master users and citizens. It enables map navigation and zoom in and out on the visually represented resource occupation. The objective is to provide user friendly map navigation for the users.
- **Resource Placement** is a functionality available for master users and citizens. Visual representation of the resources placed on the map with icons and/or tracks is provided using this functionality. The objective is to provide easy identification of resources on the map.
- **Resource Selection** is available for master users and citizens and realized by clicking on the resource icons and/or tracks represented on the tracks. This action lists the user groups occupying this resource. The objective is to provide easy resource selection and resource occupancy.
- **Simulation Visualization** functionality is available for master users and citizens to provide easy tracking of the simulation results on the map by a pie chart near resources.
- **Simulation Time Parameter Changing** is a functionality available for master users and citizens. Time change on the map for simulation representation analysis in variable

hour and day steps in order to provide easy change tracking on the resource map when time is changed.

IV. Vodno Mountain FUPOL Project Scenario

A typical scenario of the Vodno Mountain FUPOL systems' usage is presented in this Section. Simulation is based on optimization as defined by Gusev et al. [6].

The administrator creates master user and manager for the City of Skopje. The created master user logs in to the system in order to activate the simulation by uploading configuration files for both simulations.

The master user is able to start the one-year simulation (Simulation 1). It is assumed that the initial configuration is edited and stored in excel format (.xls/.xlsx) and then it is uploaded using an uploading form (import simulation configuration) on the interface. After the initial configuration is uploaded the master user is able to click on the ”Start Simulation” button. The output of the simulation contains the occupancy of each resource specified in the initial configuration. A typical example of the output is presented in Figure 5.

The master user can create, edit and change the parameters for one-week simulation (Simulation 2), using the map of the Vodno Mountain recreational center in a specified excel format (.xls/.xlsx). The configuration file can be uploaded using an uploading form. Both the master user and citizens can start the Simulation 2. They can change the parameters explained later in this document through a web interface i.e. by clicking on the map, selecting the month for the simulation and choosing the weather conditions. The simulation is started by click on the Simulate button. The output of simulation 2 is the occupancy of the resources available on the Vodno Mountain recreational center based on the user input.

For both simulations the input is sent to the simulation service. The output of the service (resource occupancy) is displayed on a map of the Vodno Mountain recreational center for easier user interaction.

The master user can also choose to export or import a specific system configuration. The export configuration button must create an excel table that can be later used for importing the configuration. Changing the configuration of the system can be done directly through the configuration files (excel tables) by changing and adding rows in the table or through a web interface where the same parameters can be changed (the parameters are discussed later in this document).

Citizens are anonymous users in the system. They can start the simulation and define their own parameters for simulation. If they would like to share their experience with the city of Skopje, or suggest a change in recreational schedule, propose to add a resource, or an activity, they can use the ticketing system i.e. create tickets with suggested schedules and/or send complains.

The manager of the system works with the ticketing system, with goal to receive all the user tickets and answer them. Reports of the ticketing system usage can be created and available to relevant users. The manager can suggest a configuration that is propagated to the master users for review.

All of the users of the system can review the simulation reports that are created from the simulation 1. Additionally the private configurations made using the simulation 2 can only be reviewed by the user that created the specific simulation. If the simulation configuration is sent than the simulation report for that configuration can be reviewed by managers and master users too.

The additional visualization functionalities for the Vodno Mountain project are:

- **Activity Placement** enables the master users and citizens to see visual representation of the activities by icons. More icons near a resource indicate increased activity on selected resource.
- **Activity Selection** is a feature realized by clicking on the activity icons. This action lists the user groups participation in the selected activity. This functionality is available for master users and citizens.
- **User Group Listing** is a visual listing of the user groups in a grid view after resource or activity selection.
- **User Group Selection** is realized by clicking on a specific row in the grid view. This action opens up the opportunity for changing the number of people in the selected user group.

V. Bike Stations FUPOL Project Scenario

This section discusses a typical scenario of the Bike FUPOL simulation project based on optimization specified by Guseva et al. [7]. Similar to the previous simulations, the administrator has the role to create master user and manager for the City of Skopje, and enable them to log in the system and activate a simulation by uploading configuration files for both simulations.

The master user can start Simulation 1 after updating the initial configuration. This contains a set of input files each predefined for a certain project that can:

- increase the quality of an existing bike path,
- build a new bike path,
- build a docking station used as bike parking, and
- build a rent-a-bike docking station.

The initial configuration consists of definitions of all bike paths, and also possible locations for docking stations to be used as bike parking lots or bike renting stations in City of Skopje. After the initial configuration is uploaded, the master user is able to click on the ”Simulate” button. A typical output is presented in Figure 6. Optimization values contain the occupancy of each parking lot or renting station specified in the initial configuration.

The master user updates the configuration files for Simulation 2, by using the map of the City of Skopje in a specified excel format (.xls/.xlsx). Both the master user and citizens can start the Simulation 2. Certain parameters can be specified prior to clicking the ”Simulate” button. The output of simulation 2 is the occupancy of the bike parking lots and renting stations in Skopje based on the user input.

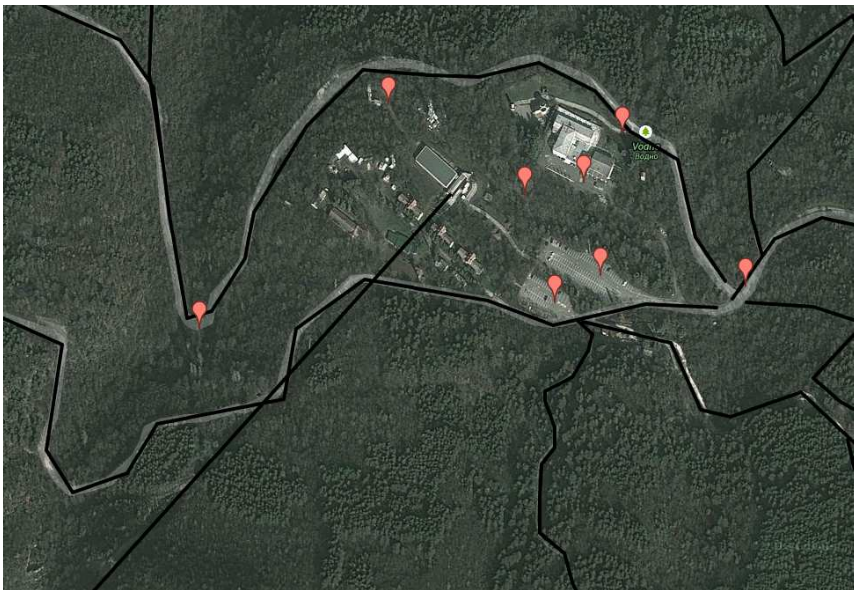


Fig. 5. An example of Vodno recreation simulator output

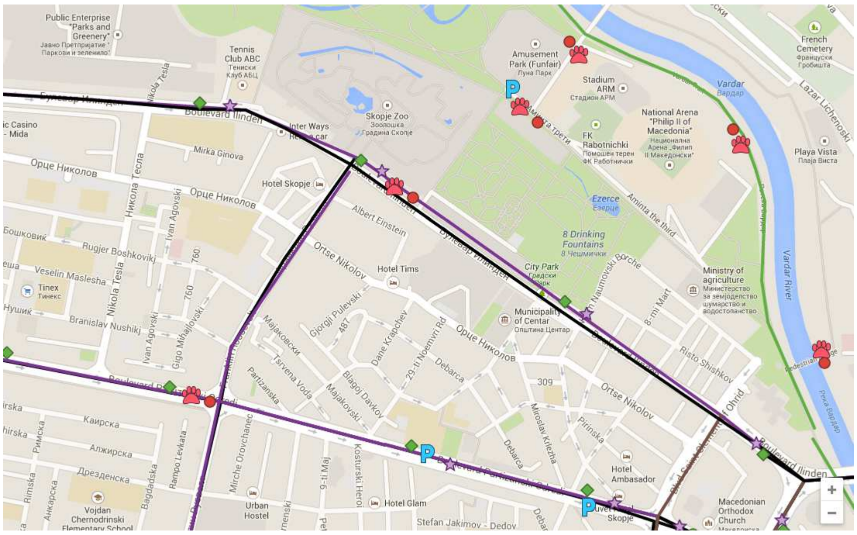


Fig. 6. An example of fostering bike inter-modality simulator output

Simulation service accepts the input files and simulates a typical behavior and calculates the resource occupancy. The output is visualized on a map.

Each input file can be exported and later imported as a specific system configuration. It automatically attaches the configuration and simulated files to a ticket a citizen is creating when trying to suggest a new project proposal. Citizens are anonymous users in the system. They can start the simulation and define their own parameters for simulation. Tickets are the means to enable citizen participate in policy creation by sharing their experience with the administration of City of Skopje.

The system manager is responsible to answer the open ticket issues or to deal with the ticketing system proposals by initializing new project proposals.

The following visualization functionalities are available for the bike simulation project:

- **Docking Station selection** is a functionality available for master users and citizens. It presents a selection of a corresponding docking station by clicking on a pointing device when pointing a corresponding icon on the map. This action presents the occupancy of the selected docking station to allow a satisfactory use of bikes as transport means. The objective of this functionality is to provide easy resource selection and resource occupancy identification.
- **Tracks selection** available for master users and citizens is a feature that is invoked by clicking on corresponding

tracks. If a track is lying under the presented one, an intermediate window is presented with a list of all tracks on a selected street part. This action also lists the occupancy of the selected tracks.

VI. Discussion

The user administration is easy to use and understand by administration of City of Skopje. So far there were no complains about complexity of social networking tools and usage of simulator tools.

All tools are realized in three languages, Macedonian, Albanian and English, making it appropriate to use by the majority of mother speaking languages and also ENglish as international language. Further multi language support is also supported. The script is also changed to enable a possibility of usage of Cyrillic and other multi language code tables.

FUPOL questionnaire and opinion map functionality has been embedded in the Blog- frontend easily and was found to be very useful for City of Skopje.

The access can be anonymous or verified by the social networking provider, such as Facebook or Twitter. Verified access is necessary to prevent spamming and other unwanted access.

The direct benefit for the citizens of the City of Skopje is that they will have a way to have their say on very important subjects, comment, criticize, and give creative suggestions, while the direct benefit for the decision makers will be an easy access to public opinion on important subjects as well as a source of ideas.

The City of Skopje in the past 2 years has created three campaigns on a variety of topics and has had a great deal of success by using the tools and software, by creating interactive maps and questioners on these topics.

Apart from the above mentioned topics the visualization includes heat maps, hot topic sensing and trend analysis which are in process of development and are still not implemented.

Additionally, two separate simulation models are being constructed, one for the recreational activities on the mountain Vodno, and the other about the planning of the bicycle track's in Skopje.

A generic approach for urban city policy modeling, including optimization and modeling is presented by Gusev et al. [8].

VII. Conclusion and future implementation

Social networking popularity is increasing and it is in the stage of mass usage. Therefore the core FUPOL platform is a set of tools that are being efficiently used by the administration of City of Skopje, to reach wider audience and enable citizen participation in policy design and modeling. The described questionnaires and opinion maps have been already subject of plenty of organized campaigns. Future developed tools like heat maps, hot topic sensing and further visualization tools will allow better use of social networks as tools for sensing the public opinion and trends.

Simulator tools are visual software packages accessed by web applications and will enable improved citizen interaction and participation in future policy modeling at City of Skopje.

Simulation 2 is mainly intended for citizens to enable them a visual tool for simulating different scenarios, based on their proposals for a new project, or reorganization of certain activities. We have to note that Simulation 1 is the most important for the administration of City of Skopje. It simulates a yearly scenario, based on input from excel tables. For example, these tables present weather conditions (for example, how many are sunny days in a particular month etc.). The simulation simulates citizen behavior as commuters according to travel needs expressed in input excel tables. For example, a citizen can decide to use a bike path if it is in corresponding quality situation marked with grade between 0 and 1, or can realize a typical recreation scenario on Vodno mountain.

Finally the output should be number of bikers, presenting the citizens using bikes as transport means in the bike project or resource occupancy in the Vodno project. Administration of Skopje will use these simulations to make a proper decision which project proposal will have the highest impact, based on certain criteria.

The output of the bike project is the number of bikers and also the capacity of a docking station. The final goal of administration of City of Skopje is to choose which project to finance based on the highest increase of number of bikers. Existing number of bikers is 2.5% of total population requesting transport, and the goal is to reach 5% average of major european cities.

References

[1] FUPOL. project web site. [Online]. Available: http://www.fupol.eu/

[2] P. Sonntagbauer, P. Boscolo, and G. Prister. Fupol: an integrated approach to participated policies in urban areas. [Online]. Available: https://www.fupol.de/sites/default/files/doc/eChallenges

[3] G. Bouchard, W. Darling, S. Clinchant, A. Mondragon, and C. Archambeau, "The fupol project: Involving citizens in policy design," in *10th International Conference on the Design of Cooperative Systems: From research to practice: Results and open challenges*, 2012.

[4] P. Sonntagbauer, A. Aizstrauts, E. Ginters, and D. Aizstrauta, "Policy simulation and e-governance," in *IADIS International Conference e-Society*, 2012.

[5] J. Kohlhammer, K. Nazemi, T. Ruppert, and D. Burkhardt, "Toward visualization in policy modeling," *IEEE Computer Graphics and Applications*, vol. 32, no. 5, pp. 0084–89, 2012.

[6] M. Gusev, B. Veselinovska, A. Guseva, and B. Gjurovikj, "Future policy modeling: A case study – optimization of recreational activities at the vodno mountain," in *Advanced ICT Integration for Governance and Policy Modeling*, P. Sonntagbauer, K. Nazemi, S. Sonntagbauer, G. Prister, and D. Burkhardt, Eds. IGI Global, 2014.

[7] A. Guseva, M. Gusev, and B. Veselinovska, "Fostering bicycle inter modality in skopje," in *Advanced ICT Integration for Governance and Policy Modeling*, P. Sonntagbauer, K. Nazemi, S. Sonntagbauer, G. Prister, and D. Burkhardt, Eds. IGI Global, 2014.

[8] M. Gusev, G. Velkoski, A. Guseva, and S. Ristov, "Urban policy modelling: A generic approach," in *ICT Innovations 2014, AISC 311*. Springer, 2015, pp. 187–196.

Open Public Transport Data in Macedonia

Elena Mishevska, Bojan Najdenov, Milos Jovanovik, Dimitar Trajanov
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Republic of Macedonia

Abstract—**The need to represent data on the Web in a way that will make it easier to manage, has led to new solutions for data representation, visualization, storage and querying. The concepts of Open Data, Linked Data and the Semantic Web offer a significant improvement in information and data dissemination. These concepts aim towards making data on the Web machine-readable and enable interlinking between data from different datasets, published on different locations. This allows easier data retrieval by software agents, and enables use-case scenarios which are unavailable over isolated data silos. On the other hand, personal time management and daily commute navigation in urban areas are one of the biggest influencers on the quality of life of a person. Public transport data has high value for citizens and generates numerous use-cases. In this paper, we describe the process of obtaining data from the public transport company JSP Skopje, transforming them into the standardized Google Transit Feed Specification[1] format, enhancing them and creating 4 star Open Data. We reused the Transit Ontology[2] and the W3C Geospatial Vocabulary[3], and developed our own complementing ontology for annotation purposes. We published the generated RDF datasets in order to support the provided use-case scenarios from this domain via a public SPARQL endpoint.**

Keywords—Public Transport; Open Data; GTFS; RDF; Ontologies;

I. INTRODUCTION

The Open Data concept represents the idea that data generated by the governments, their institutions and other public entities, and which is public by its nature, should be published in an open, raw and machine-readable format. This data can then be used, reused, republished and redistributed, generally in applications which leverage the value of the data [1]. This allows for a variety of useful applications to be built with the published datasets, using and combining data in various ways. Further linking of this data with data from other datasets, extends the possibilities, by providing the opportunity to use data and information relevant for the use-case, but not part of the original dataset. Linked Data is about employing the Resource Description Framework (RDF) and the Hypertext Transfer Protocol (HTTP) to publish structured data on the Web and to connect data between different data sources, thus effectively allowing data from one dataset to be interlinked with data in another dataset [2].

Public transport is a very important issue in the lives of citizens. Therefore, public transport data caries values which are of great importance for the levels of quality of life. This was our main motivation to work with public transport datasets and try to create data representations which provide various use-case scenarios via REST services.

Another thing which motivated us to work on this subject was the Helsinki Open Transport Data Manifesto[4], which empowers the free flow of transport data across Europe and puts focus on the opportunities and benefits of opening up and sharing this reach resource to the stakeholders.

A. The Star Rating System for Data

In order to classify the data published on the Web by its availability and its usefulness, there is a standard star rating system. According to the rating system[5], every information which has been published online can be considered as Open Data, and is given a one star rating. This usually includes images, PDF documents, and other documents types as well.

Making the data machine readable earns it a two star rating. This usually includes Microsoft Excel spreadsheets. Publishing data in non-proprietary format, such as CSV, earns it a three star rating. Published Open Data datasets, which use Semantic Web standards (RDF, RDFS, OWL) for annotation of the entities, earn four stars. The last one, the fifth star, is given for datasets which contain links towards other, already published dataset on the Web, in order to provide context [3].

Almost all of the Google Transit data published in the GTFS format are Open Data by these definitions, since they are available online for public use. However, since the GTFS format of the data obligates data to be represented in strictly formed CSV files (explained in more details further in the paper), no effort has been made to bring public transportation data up in the star rating system.

[1] https://developers.google.com/transit/gtfs/reference
[2] http://vocab.org/transit/terms/
[3] http://www.w3.org/2005/Incubator/geo/XGR-geo/
[4] http://www.epsiplatform.eu/transport
[5] http://5stardata.info/

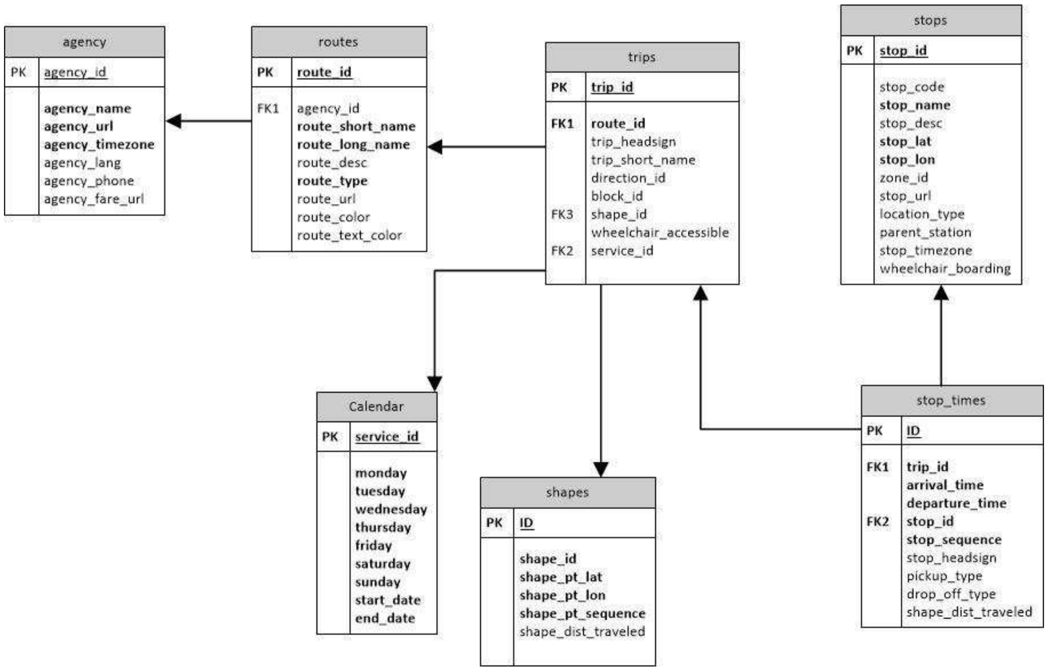


Figure 1. The GTFS Schema for the data from JSP Skopje.

We decided to leverage local public transportation data from Macedonia as high as possible in the rating system, by reusing existing ontologies and combining them with our own ontology for the annotation purposes. We used data collected from the JSP Skopje website, which has been transformed by another project into GTFS format in order to be used with the Google Transit system. The dataset is described in more details further in the paper.

II. Related Work

The Google Transit system is widely used around the globe from many transport companies from the six continents that use different kinds of transportation. Google Transit is available for companies from USA, Austria, Belgium, Australia, Ghana, Nigeria, Bulgaria, etc[6]. However, very little effort has been made to semantically annotate any of the Google Transit data so far, or any other transit data, not necessarily in the GTFS format.

Data from the New York Subway in GTFS format have been published as RDF files, but the files, as they are presented, have not been created using any ontologies and are not linked anywhere on the Web[7]. Also, a tool for transforming GTFS data into RDF exists. The tool is written in Perl, and uses a Turtle[8] syntax to map the RDF files. An Android application, called GetThere[9], which uses Linked Open Data has been developed in order to provide information about the location of busses in rural areas in the UK [4].

As for transport open and linked data, there are only a few datasets from Great Britain and France [5], but they concentrate on the physical locations of the stops or the connectivity of the different stops or cities. For now, no linked data datasets of any transit agency are available on the LOD cloud[10].

III. Google Transit Feed Specification

The data consists of multiple CSV format files (with the .txt extension), each representing a different piece of data, written in a strictly specified form. Not all of the schema tables are obligatory, and not all of the schema tables exist for different publishers [6].

JSP Skopje[11] is a public transportation company from Skopje, which provides public transit services on the territory of Macedonia's capital. The transport is done by busses which operate on different routes following predefined schedules. JSP Skopje publishes the bus schedule on their website in standard HTML format.

As part of another project at our Faculty, the data from the JSP Skopje website have been collected and transformed into the GTFS format, for the purposes of them being used as part of the Google Transit system. The JSP Skopje GTFS dataset consists of seven of the standard GTFS tables (Figure 1):

- agency - contains information about one or more transit agencies.
- stops - contains information about all the stops.
- routes - contains information about the routes of the transit agency.

[6] http://maps.google.com/landing/transit/cities/index.html
[7] http://www.cs.sunysb.edu/~pfodor/new_york_subway_data/
[8] http://www.w3.org/TR/turtle/
[9] http://gettherebus.com/
[10] http://lod-cloud.net/
[11] http://jsp.com.mk/

- trips - contains sequences of two or more stops that occur at a specific time.
- stop_times - contains vehicle arrival and departure times from individual stops for each trip.
- calendar – contains schedule information.
- shapes - the spatial representation of a route alignment so it can be accurately drawn on a map [7]

IV. Transforming the GTFS Data From JSP Skopje into Four Star Open Data

Besides having the transport data from JSP Skopje in GTFS format, we believe that creating a semantic annotation of the data and publishing it as four star Open Data on a SPARQL endpoint, will make it easier to use. This endpoint would serve as a REST service for developers to effectively use from their applications, built over the dataset and providing public transit system services for the city of Skopje.

In order to start the annotation process, we needed a suitable ontology.

A. Ontology

The most common approach and the best practice in providing an ontology, is reusing an already developed one. Additionally, one usually has to creating his own ontology for the properties which do not exist in other ontologies, in order to start with the annotation process.

Our search for a suitable ontology lead us to the Transit Ontology[12], developed for similar dataset to the one we had, a Google Transit dataset, but developed for a different scenario which was a bit different than ours. Regardless, the ontology provided enough properties we could match with our data. The ontology provided us with some of the classes and the properties we needed, and what was left for us was to find a solution for the remaining properties. Therefore, we reused some properties from the W3C Geospatial Ontology[13], and we created an ontology which covered the remaining properties.

Here, we provide a listing of the classes and properties we used from the Transit Ontology. The diagram of the ontology is shown on Figure 2.

The Transit Ontology we reused provided us the following classes:

- Transit Route – a public transportation route.
- Transit Stop – a location where passengers board or disembark from a transit vehicle.
- Transit Agency – an organization that oversees public transportation for a city or region.

The Object properties we used to link data from the previous classes were:

- Route – the route the trip of interest uses;
- Stop – a physical location connected to a service stop;
- Agency – the agency that operates the route of interest.

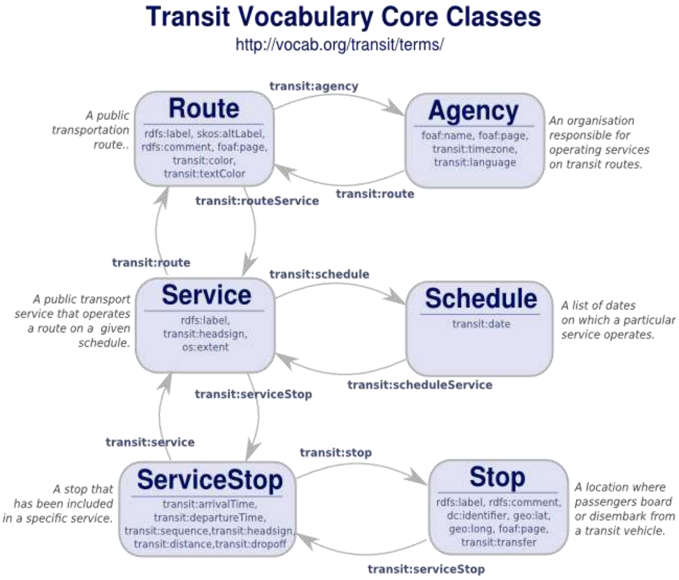

Transit Vocabulary Core Classes
http://vocab.org/transit/terms/
Route
rdfs:label, skos:altLabel, rdfs:comment, foaf:page, transit:color, transit:textColor
A public transportation route..
Agency
foaf:name, foaf:page, transit:timezone, transit:language
An organisation responsible for operating services on transit routes.
transit:agency
transit:route
transit:routeService
transit:route
Service
rdfs:label, transit:headsign, os:extent
A public transport service that operates a route on a given schedule.
Schedule
transit:date
A list of dates on which a particular service operates.
transit:schedule
transit:scheduleService
transit:serviceStop
transit:service
ServiceStop
transit:arrivalTime, transit:departureTime, transit:sequence,transit:headsign, transit:distance,transit:dropoff
A stop that has been included in a specific service.
Stop
rdfs:label, rdfs:comment, dc:identifier, geo:lat, geo:long, foaf:page, transit:transfer
A location where passengers board or disembark from a transit vehicle.
transit:stop
transit:serviceStop

Figure 2. Transit Ontology Diagram.

The DataType properties we reused were:

- Timezone - the time zone where a person or organization is located.
- Language - the primary language used by a person or organization.
- Sequence - a sequence number for a stop along a route or service.
- Distance - the distance of this service stop from the first stop in sequence.
- Headsign - text that appears on a sign that identifies the service's destination to passengers.
- Color - a color associated with this route.
- Text Color - a legible color for text drawn against a background of the color associated with a route.

For the latitude and longitude data for the location of the stops and the locations of the different stop points, we used the ‘latitude’ and ‘longitude’ properties of the W3C Geospatial Ontology.

We defined the rest of the classes and properties in our own ontology, GTFS-ext, which extends the Transit Ontology and provides us with all of the classes and properties we needed in order to fully map the available GTFS data from JSP Skopje.

The Classes we added to our ontology are given in Table 1. The Object properties that we defined in order to link the classes can be seen in Table 2. Additionally, we created all the necessary Datatype properties which were not defined in the Transit Ontology, but were necessary for annotating our dataset, according to the GTFS Schema from Figure 1.

[12] http://vocab.org/transit/terms/
[13] http://www.w3.org/2005/Incubator/geo/XGR-geo-ont/

Table 1. Classes introduced in our GTFS-ext Ontology.

Class	Description
Shape	Specification about how the lines are represented on the map
Trip	Sequence of two or more stops that occur at a specific time
stop_times	Arrival and departure times from an individual stop
calendar	Provides schedule for a specific service

Table 2. Object Properties from our GTFS-ext Ontology.

Object Property	Description
service	References the service of a specified trip
st_times	Connects a trip with the specific departure and arrival times
shape	Shape associated with the referenced trip

B. Mapping the Data from CSV to RDF

The next step was mapping the data and transforming it from CSV to RDF. In order to achieve this, we used a Virtuoso Universal Server[14] instance, which provides mechanisms for transformation and management of various types of data. It serves as a Linked Data server and allows local and remote data querying with the Semantic Web query language, SPARQL. This is enabled over a public SPARQL endpoint which can be used as a REST service.

The mapping process was conducted into several stages. First, we imported the CSV files into relational databases in Virtuoso. Then, using R2RML[15] – a mapping language for transforming RDB data into RDF data – we created RDF Views over our relational databases.

The R2RML mapping was done using mapping files which contain information about the transformation of the RDB tables into RDF triples, each contains a subject, a predicate and an object. Each row of the relational database represents a unique entity, and each column a new triple with the entity as a subject. We used the previously discussed ontologies to annotate the data. Most of the entities from the tables are identified with the identifiers which were part of the GTFS CSV data. A small portion of the data (e.g. stop_times) is identified by the row number of the input file.

After all of the data was mapped, we ended up having seven individual graphs, one for each of the tables (Figure 1) containing the data. The next thing we needed to do was to link the graphs, i.e. create the links that connected the different pieces of information into usable information. RDF links take the form of RDF triples, where the subject of the triple is a URI reference in the namespace of one dataset, while the object of the triple is a URI reference in the other [8]. We achieved that using the SPARQL endpoint from the Virtuoso server and by using the SPARQL query language we combined and created the appropriate links for the different graphs. We created the following links:

- Linked the ‘routes’ graph with the ‘agency’ graph using the ‘agency’ object property.
- Linked the ‘trips’ graph with:
  - the ‘routes’ graph, using the ‘route’ property.
  - the ‘calendar’ graph, using the ‘service’ property.
  - the ‘shape’ graph, using the ‘shape’ property.
  - the ‘stop_times’ graph, using the ‘st_times’ property.
- Linked the ‘stop_times’ graph with the ‘stops’ graph using the ‘stop’ object property.

With that, we finished the process of semantic annotation of the JSP Skopje transit data.

The idea of linking the resulting four star RDF dataset with another similar datasets, in order to provide a wider range of additional use-cases, led to no success. We were unable to find other semantically annotated transit datasets related to ours in any way, which would make sense to interconnect.

V. Use-Cases

The basic idea behind the semantic annotation and leveraging the public transport data from JSP Skopje to four star Open Data, is creating a publicly available dataset which could be easily used and would provide a large variety of use-case scenarios involving public transport.

With our transformed dataset, we can provide information about the stop and the time the travelling party should go to, in order to get a bus that could take them to the desired location within the city of Skopje, what is the stop they need to get off, as well as the arrival time on that stop. This all would represent a search in our graph to find the departures on a specific stop, at a specific time which will match a route that suits our needs.

A. Use-Case 1

If, for example we want to find the departure times of a given bus route in a given period of time during a specific time of the year, we could generate a use-case similar to the following: we look for departure times of the route ‘R15’, from the ‘Ново Лисиче’ station, running between 09:00 – 10:00 AM, during weekdays in winter. For this, we can use the following SPARQL query:

```
prefix ont:
<http://linkeddata.finki.ukim.mk/lod/ontology/transit-
ont#>
prefix transit: <http://vocab.org/transit/terms/>
select distinct ?departure
where {
        graph
        <http://linkeddata.finki.ukim.mk/lod/data/routes#>
        { ?r ont:route_id "R15" }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/calendar#>
        { ?s ont:service_id "DELNIK_ZIMEN" }
```

[14] http://virtuoso.openlinksw.com/

[15] http://www.w3.org/TR/r2rml/

```
        graph
<http://linkeddata.finki.ukim.mk/lod/data/trips#>
        { ?x transit:route ?r ; ont:service ?s ;
             transit:headsign "Карпош 4" ;
             ont:st_times ?st.
        }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stop_times#>
        { ?st ont:departure_time ?dep;
              transit:sequence 1. }
        FILTER regex(?dep,"(^09:)|(^9:)")
}
```

The results from the query are shown in Table 3.

This query obtains the URI of the route with the ‘route_id’ equal to ‘R15’ from the ‘routes’ graph and the URI of the service with the ‘service_id’ equal to ‘DELNIK_ZIMEN’ from the ‘calendar’ graph. Then, from the ‘trips’ graph, it finds the trips which correspond to the route and schedule we previously obtained and which have the ‘transit:headsign’ property set to ‘Карпош 4’ (the direction of the bus). After that, it finds the records from the ‘stop_times’ graph which correspond to the selected trips using the ‘tr:st_times’ property. Finally, it filters the properties that fulfill the condition that the departure time is between 09:00 - 10:00 AM.

Table 3. Results from the SPARQL query for Use-Case 1.

DEPARTURE
"09:55:00"

B. Use-Case 2

In another use-case scenario, we may want to find all the routes that go through a specific bus stop. This will require the usage of four graphs: the ‘stops’ graph, to find the stop URI, the ‘stop_times’ graph to find arrivals and departures from the specific stop and the trips that use that stop, the ‘trips’ graph to find the routes associated with the specified trip, and finally, the ‘routes’ graph to find the route’s name. Let the stop we use be a more frequent one, for example the “МАЛ ОДМОР” stop.

To select the routes which go through the “МАЛ ОДМОР” stop along with their names, we use the following SPARQL query:

```
prefix transit: <http://vocab.org/transit/terms/>
prefix ont:
<http://linkeddata.finki.ukim.mk/lod/ontology/transit-
ont#>
select distinct ?r ?n
where {
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stops#>
        { ?o ont:name "МАЛ ОДМОР" }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stop_times#>
        { ?s transit:stop ?o. }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/trips#>
        { ?t ont:st_times ?s ;
            transit:route ?r.
        }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/routes#>
        { ?r ont:name ?n }
}
```

The results from this query are shown in Table 4.

We will see how the SPARQL query achieves the result step by step. First, starting from the ‘stops’ graph, using the ‘ont:name’ property, it selects the appropriate URI of the stop and uses it in the ‘stop_times’ with the ‘transit:stop’ property to select the arrivals and departures on the selected stop. The ‘transit:stop’ property is a property of the transit ontology we reused. After that, the query passes the URIs of the arrival and the departure values to the ‘trips’ graph, in order to find the trips that arrive at certain times at the stop, using the ‘ont:st_times’ property and using the ‘transit:route’ properties selects the routes associated to the trips selected with the ‘ont:st_times’ property. Finally, in the ‘routes’ graph, the query uses the ‘ont:name’ property to find the already selected routes’ names and presents the final results.

Table 4. Results from the SPARQL query for Use-Case 2.

R	**N**
http://vocab.org/transit/terms/ Route/R4	"11 Октомври - Нас. Хром"
http://vocab.org/transit/terms/ Route/R2	"Сарај - Автокоманда"
http://vocab.org/transit/terms/ Route/R7	"Нас. Лисиче - Карпош 3"
http://vocab.org/transit/terms/ Route/R15	"Ново Лисиче - Карпош 4"
http://vocab.org/transit/terms/ Route/R22	"Транспортен центар - Волково"
http://vocab.org/transit/terms/ Route/R59	"Карпош 3 - Гробишта Бутел"
http://vocab.org/transit/terms/ Route/R19	"Шуто Оризари - Карпош 4"
http://vocab.org/transit/terms/ Route/R24	"Кисела Вода - Тафталиџе"

C. Use-Case 3

We can have a scenario in which we want to count the number of bus trips which occur at a selected bus stop, during a specific schedule. For example, we can take ‘NEDELA_ZIMEN’ as a schedule, and ‘ПАЛМА’ as a bus stop. This way, we will count the number of buses trips which will pass at this bus stop, during Sundays in winter.

The SPARQL query for this scenario is:

```
prefix ont:
<http://linkeddata.finki.ukim.mk/lod/ontology/transit-
ont#>
prefix transit: <http://vocab.org/transit/terms/>
select count(distinct ?t) as ?count
where {
        graph
<http://linkeddata.finki.ukim.mk/lod/data/calendar#>
        { ?s ont:service_id 'NEDELA_ZIMEN' }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/trips#>
        { ?t ont:service ?s ; ont:st_times ?st . }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stop_times#>
        { ?st transit:stop ?stop }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stops#>
        { ?stop ont:name "ПАЛМА" }
}
```

The result from the query is shown in Table 5.

Table 5. Result from the SPARQL query for Use-Case 3.

COUNT
193

We could run a similar query, but for the Saturday winter bus schedule, and compare the results with the one in Table 5. The SPARQL query for this would only require us to the schedule (‘service_id’ property) to “SABOTA_ZIMEN”. The result of this query is Count: 275, which leads us to the conclusion that public transport bus lines in Skopje are more frequent on Saturdays than Sundays, which is expected.

D. Public Data Endpoint

It is also important to notice that these SPARQL queries can be sent to a SPARQL endpoint[16] on our live Virtuoso instance, in a REST service manner. This means that applications which would potentially use this dataset can query the data with simple HTTP GET requests and obtain the data in a variety of RDF and non-RDF formats, such as RDF/XML, Turtle, JSON-LD, RDF/JSON, N3, JSON, CSV, HTML, etc.

The general format of the HTTP calls is:

```
http://linkeddata.finki.ukim.mk/sparql?
query=SPARQLQUERY&format=FORMAT
```

VI. CONCLUSION AND FUTURE WORK

The Open Data concept holds the key to organizing and collecting information from the public sector, and using it in various ways and use-case scenarios. By publishing and linking data in the right way, we can gain more useful information and extract properties which do not even have to exist in our dataset, but can carry information of enormous relevance. And today, data represents the new oil for the industry [9].

In a similar manner, publishing and interlinking transportation data can take public transportation, trip planning, even sightseeing to a whole new level.

In the paper we gave an overview of the process of transforming the public transport data from JSP Skopje, into four star Open Data. We worked with big datasets, and created RDF graphs which contain 6.005.619 triples, out of which 5.227.956 are from the ‘stop_times’ graph, 725.758 are from ‘trips’, 46.914 are from ‘shapes’, 4.649 are from ‘stops’, 170 are from ‘routes’, 165 are from ‘calendar’ and only 7 are from the ‘agency’ graph.

We also provided example use-cases, in hope to encourage multiple stakeholders to start publishing Open Data from the public transportation sector and also, to start thinking about creative ways to use the available data. The annotated data and the use-case scenarios are accessible via HTTP GET requests to our public SPARQL endpoint.

In the future, we would continue our work in this sector and semantically annotate datasets from more transit agencies in Macedonia, both intercity and international ones. We would interlink them to provide more useful use-case scenarios which would allow trip planning throughout the country and would create an opportunity for developing application which could implement the features provided by the interlinked datasets. Hopefully, this will raise the awareness of the possibilities the usage of Open Data provides, and especially Open Transport Data.

On the other hand, we also hope to encourage more transit agencies, not only from Macedonia, but also from all around the world, to publish semantic annotated data, which will allow interlinking the data, leading to better usage of the published data and allowing the development of even more powerful solutions and more useful applications.

ACKNOWLEDGMENT

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje, as part of the research project “Semantic Sky 2.0: Enterprise Knowledge Management”.

REFERENCES

[1] T. Berners-Lee, N. Shadbolt, “There’s gold to be mined from all our data”, The Times, 2012.

[2] C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, ”Linked Data on the Web”, LDOW, 2008.

[3] M. Jovanovik, B. Najdenov, D. Trajanov, “Linked Open Drug Data from the Health Insurance Fund of Macedonia”, 10th Conference for Informatics and Information Technology (CIIT), 2013.

[4] D. Corsar, P. Edwards, C. Baillie, M. Markovic, K. Papangelis, J. Nelson, “GetThere: A Rural Passenger Information System Utilising Linked Data & Citizen Sensing”, International Semantic Web Conference (ISWC), 2013.

[5] J. Plu, F. Scharffe, “Publishing and linking transport data on the Web”, First International Workshop On Open Data (WOD), 2012.

[6] N. Kizoom, P. Miller. A Transmodel based XML schema for the Google Transit Feed Specification With a GTFS / Transmodel comparison. Crown, 2008.

[7] A. Antrim, S. J. Barbeau, “The Many Uses of GTFS Data – Opening the Door to Transit and Multimodal Applications”, ITS World Congress (ITSWC), 2013.

[8] C. Bizer, T. Heath, T. Berners-Lee, “Linked Data - The Story So Far”, Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, 2009.

[9] V. Kundra, “Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect”, Joan Shorenstein Center on the Press, Politics and Public Policy, 2012.

[16] http://linkeddata.finki.ukim.mk/sparql

Session 6

eWorld – eWork, eCommerce, eBusiness, eLearning 2

Optimization Models for Future Policy Implementation: A Case Study

Biljana Veselinovska, Ana Guseva
Innovation Dooel
1000 Skopje, Macedonia
{biljana.veselinovska, ana.guseva}@innovation.com.mk

Marjan Gusev
Ss. Cyril and Methodius University
1000 Skopje, Macedonia
marjan.gushev@finki.ukim.mk

***Abstract*—This paper presents optimization models intended to support the policy design and implementation life-cycle. The innovations are driven by the demand of citizens and political decision makers to support the policy domains in urban regions with appropriate ICT technologies. It will target domains such as sustainably development, urban planning, and more specifically, the scheduling of different recreational activities at the Vodno Mountain, as well as fostering bicycle inter-modality in the city of Skopje. Optimization functions are defined as an input to a simulation model, which will be developed in the future. Also, citizens are included in the process of policy decision making by creating social media surveys and gathering online public opinion. Thus, citizens would also understand how and why certain decisions and laws will be imposed.**

Index Terms*—*FUPOL*; *Optimization*; *Policy implementation*; *Computer simulation

I. Introduction

Governance is related to rules, processes, and behaviours by which interests are articulated, resources are managed, and power is exercised. All these activities are realised in a society by the institutions in charge of public affairs. Participation of all relevant actors and citizen is allowed and highly appreciated. Centrally-led governing development policies alone cannot address the complexities of sustainable development successfully. Local Authorities' political recognition has not always been accompanied by an adequate level of autonomy, capacity development and financial resources. Given this fact, leaves their empowerment incomplete and a generates a certain number of obstacles. They have to face these challenges to unlock their development potential.

This potential is primarily linked to the way Local Authorities manage and implement public policies and services on the basis of local policy-making processes. Very important are the interactions with other public institutions, citizens and private sector. Thus, it is fundamental to simultaneously empower local public authorities and citizens, to ensure both that the latter have the ability to demand transparency and accountability, and that Local Authorities have the means and incentives to respond to citizen demands for effective, transparent and accountable governance, as well as an equitable allocation of resources and access to services.

This is particularly so at the local level, where citizens live and work, where basic services are provided and where enterprises are established. Citizens have, therefore, common interests at stake, to set objectives and work together in identifying solutions particularly aiming at improved access to services, a more balanced distribution of available resources, greater social cohesion and enhanced accountability and transparency of public authorities, including to accountability mechanisms.

Being closer to citizens than other public institutions, Local Authorities hold responsibility in mobilising local societies' opinions. This is particularly true in terms of more efficient public administration, more inclusive development processes, in cooperation with Civil Society Organisations' (CSO-s), and solutions to urgent challenges faced by local communities.

Local authority challenges include social exclusion, migration, food security, limited infrastructures, rapid urbanisation, depletion of resources, public safety and violence, environmental and social impact of extractive activities, climate adaptation and mitigation, rule of law and access to justice. Furthermore, the guidance of Local Authorities and the mobilisation of additional private and community capacities and resources can trigger a change in the quality of citizens' life and wellbeing, ensuring a balance between socio-economic growth, equity and environmental quality and increasing the resilience of the most vulnerable.

Therefore, the FUPOL project has been introduced. This project proposes a comprehensive new governance model to support the policy design and implementation life-cycle. The innovations are driven by the demand of citizens and political decision makers to support the policy domains in urban regions with appropriate ICT technologies. It specifically targets domains such as sustainable development, land use, urban planning, urban segregation and migration. The aim is reducing the complexity through a comprehensive policy spiral design life-cycle approach deemed appropriate for complex societal problems.

The outcomes of the project, designed in line with the ICT work program, include a new governance model to engage all stakeholders in the whole policy design life-cycle, a policy knowledge database, a cloud computing based comprehensive ICT Framework, multilingual training, piloting in Europe and China, large scale dissemination and a sustainable exploitation strategy.

The FUPOL framework to support the policy lifecycle contains major innovations, namely multichannel social computing, policy topic sensing and extraction, advanced visualization including integration with GIS, multilingual semantic analysis, advanced policy modelling and model repository, dynamic agent based simulation, cloud computing and IMS supported crowd sourcing. Furthermore, FUPOL is expected to lead to better policy decisions, more efficient implementation of government policies as well as better identification of consequences for citizens and businesses.

In accordance to FUPOL's main objectives, our research focuses on improving policy decisions, more efficient implementation of government policies as well as better identification of consequences for citizens and businesses through developing optimization models which will provide an input for a simulation tool, to be created in the future. Prior to our optimization models, social media tools will be used to gather information about the citizens preferences concerning the policy decision making process. By improving the policy decision making process, this research will achieve optimal scheduling of recreational activities at the Vodno mountain, as well as encourage the growth of bicycle inter-modality in Skopje, and most importantly, include citizens opinions about these specific problems.

The rest of the paper is organized as follows. The motivation for this research in the area of optimization is presented in Section II. Sections III and IV describe the newly defined optimization models for scheduling of recreational Activities at Vodno Mountain, and fostering Bicycle Inter-modality in Skopje respectively. Section **??** state the benefits and impacts of the implementation of FUPOL. Finally, the conclusion is specified in Section VI.

II. Motivation

Firstly, we will explain the term optimization and its characteristics. Pinter [1] states that optimization concepts and tools are frequently used in the process of making engineering, economic and scientific studies quantitative decisions. Optimizing means finding the "absolutely best" decision, which is illustrated by the minimum (or maximum) of a suitable objective function, while it satisfies a given collection of feasibility constraints. This objective function expresses overall (modeled) system performance, such as loss, risk or error. The constraints include physical, technical or other considerations [1].

These formal descriptions of the optimization process helped us in defining the optimization functions concerning the input for recreation activities schedule in Vodno, as well as explore the barriers and facilitators in using bicycles as a transport mean in Skopje, which will be explained later in this paper. This research encourages the scheduling of recreational activities in the forest area of Vodno, and fostering bicycle inter-modality is Skopje, which is an essential necessity to all the citizens on a daily basis. Thus, we can define four basic values, the goal of which is to improve the quality of life, constitute the "nucleus" of the idea for a rational and healthy development of the city: development, environmental sustainability, freedom, as well as solidarity.

We can define four basic values, the goal of which is to improve the quality of life, constitute the nucleus of the idea for a rational and healthy development of the city:

1) **Development.** The concept of development goes beyond the economic prosperity or growth. It includes also social, cultural, traffic and other aspects.
2) **Environmental Sustainability.** Development must not deprive the future generations of their chances. Sustainability in this sense is not achieved everywhere. However, it is possible to arrive even closer to this value with the help of well-planned and properly oriented actions.
3) **Freedom.** Freedom implies the possibility of making choices in accordance with individual preferences within the limits defined with respect to those other people wishing the same. There is no freedom without participation. Dependence is an important means for achieving participation. It also helps to improve the quality of decisions.
4) **Solidarity.** Solidarity is a characteristic of a society that cares and that shares the benefits of development. Furthermore, we have defined and explained our research's objectives which are listed below:
   - Encouraging the cooperation between the stakeholders at all levels, in the interest of environment improvement;
   - Developing and improving the existing infrastructure in the forest areas;
   - Encouraging protection, conservation and regeneration of the environment;
   - Improving the recreational activities infrastructure;
   - Supporting and creating conditions for investments, especially for the marketing; Improving the cooperation in the field of health care, education etc. [2]

III. Optimization Model: Recreational Activities at Vodno Mountain

The first case study is about optimization model that aims at organizing the recreational activities at Vodno Mountain while preserving its natural environment and avoiding conflicts as much as possible. The output is an optimized schedule of different recreational activities on the Vodno Mountain, satisfying the defined constraints, making it an optimisation problem to find a minimal function and satisfying the constraints as a system of inequalities.

Although the minimisation of conflicts may be also considered as a traffic related problem, here we express different perspectives and objectives in solving the problem. The overall goal is to find a schedule of recreation activities that enables optimal (minimal) values of analysed function for (contamination) pollution and noise pollution, maximal values of natural preservation and cleanliness, criminal and physical safety.

Assume that the number of citizens N using Vodno recreation facilities with corresponding activity A_i, for $i = 1, 2, \ldots, N_A$ on a specific resource R_j, for $j = 1, 2, \ldots, N_R$ in an one hour time slot t_k, where $k = 0, 1, 2, \ldots, 23$ is denoted by $N_{ij}(t_k)$.

The overall optimisation function is summarised as follows: Find a schedule that identifies $N_{ij}(t_k)$ by optimising

- *minimal pollution function* (PL), obtained as a function that calculates (1) as sum of all those performing the activity A_i on a resource R_j adjusted by a correlated pollution coefficient C_{PL}.

$$PL = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{PL}(A_i, R_j) N_{ij} \quad (1)$$

- *minimal noise pollution function* (NP) calculated by (2) as sum of those that realise recreation activity A_i on a resource R_j weighted by the correlated noise pollution index C_{NP}.

$$NP = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{NP}(A_i, R_j) N_{ij} \quad (2)$$

- *minimal number of conflicts* (FN) for a given time slot t_k, calculated by (4) as a sum of conflicts and appropriate number of those using the conflicted activities on a given resource. A conflict is defined by (3) performing two conflicting activities A_i and A_l on a given resource R_j. The coefficient CN_{ij} can have value 1 if there is a connection, meaning that the activity A_i is allowed on the resource R_j and 0 if there is no connection, i.e. the activity is not allowed. The index CM also gets a value of 1 if both activities are conflicting (such as the activity of walking with children and the activity biking on the same resource), and otherwise 0.

$$FN_{i,j,l} = CN(A_i, R_j) CN(A_l, R_j) CM(A_i, A_l) \quad (3)$$

$$FN(t_k) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} \sum_{l=1}^{N_A} FN_{i,j,l} N_{ij}(t_k) N_{lk}(t_k) \quad (4)$$

- *maximising the criminal and physical safety*, calculated correspondingly by (5) and (6). The coefficients C_{CS} and C_{PS} depend on provision of an activity A_i on a resource R_j, and may get values in the range between 0 and 1. The higher the value is, means the higher safety.

$$CS = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{CS}(A_i, R_j) N_{ij} \quad (5)$$

$$PS = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{PS}(A_i, R_j) N_{ij} \quad (6)$$

- *maximising the natural preservation and cleanliness*, calculated by (7) and (8). The coefficients C_{NR} and C_{CL} obtain the values in the range between 0 and 1 and depend on activity A_i performed on the resource R_j.

$$NR = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{NR}(A_i, R_j) N_{ij} \quad (7)$$

$$CL = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{CL}(A_i, R_j) N_{ij} \quad (8)$$

The overall optimisation schedule is due to several constraints, such as the following

- *coefficient range*, as expressed in (9):

$$C_{PL}, C_{NP}, C_{CS}, C_{PS}, C_{NR}, C_{CL} \in [0, 1] \quad (9)$$

- *maximum capacity*, constraints of reaching full capacity of a given resource as expressed in (10)

$$N_{ijmin}(t_k) \leq N_{ij}(t_k) \leq N_{ijmax}(t_k) \quad (10)$$

A detailed elaboration about coefficient and the model can be found in [3].

The straight forward solution is by solving the set of inequalities, but this approach is compute-intensive and time demanding. Most of the realisations use greedy algorithms that eliminate the computation space.

IV. Optimization Model for Bike Use

Bike use is a transport problem and can be analysed by conventional methods that analyse the traffic and try to find an appropriate schedule that will eliminate the conflicts as much as possible. However, in this paper we analyse it with a different perspective, trying to increase the overall number of those that will choose bike as transport means and at same time to decrease the number of those using cars and public transport. The measures and activities reaching these objectives will allow a more healthier environment for the citizens of Skopje [2]. The problem initiated in City of Skopje as a case study, are common to all other cities and can be achieved in at least the following projects for:

- *improving the existing bike infrastructure*, which include activities to increase the quality of bike existing bike paths by repairing the holes, eliminating the road obstacles, etc.,
- *developing a new bike infrastructure*, consisting of measures to build new bike paths with high quality,
- *introducing bike docking stations* aiming to build bike renting and bike parking facilities,
- *fostering intermodality* with goal to enable conditions where the commuters will use alternative transport means in the daily transportation of people and goods, more specifically, including the use of bicycles, besides usage of public transport, or car parking and rent-a-bike facilities.

Latest data shows that bike use in Skopje [4], [5], [6] varies between 1.4% to 2,5%, in comparison to 5% to average European cities. This case study analyses the overall goal of administration of City of Skopje to increase the percentage of bikers in comparison to the total number of commuters. It is in essence an optimisation problem, since there is a list of projects that can be proposed, and within the limited budget the administration should choose only a subset that fits in the budget and impacts the most.

Lets assume that there are $N_{projects}$ project proposals in a set S. The idea is to find a subset S_{opt} which consist of smaller number of projects and satisfies the constraints and limitations and in the same time results in minimal pollution and maximum number of bikers. Since the problem of choosing a subset of a given set is exponentially growing with degree of $2^{N_{projects}}$ this problem in its essence is exposes exponential processing demands. The formal specification of the optimising function is specified in (**??**).

Find a subset of projects $S_{opt} \subseteq S$ by optimising the following:

- *minimal average transport time (T_{avg})*, obtained as a function that calculates the average transport time by using a public transport between a set of predefined city locations in time intervals exposed to the most traffic jams (morning and afternoon rush time); by calculating (11), where M is the number of predefined trips among selected locations and corresponding measured time is t_i, for an integer parameter i in the interval $[1 \ldots M]$

$$T_{avg} = \frac{\sum_{i=1}^{M} t_i}{M} \tag{11}$$

- *minimal pollution function P*, where the pollution function is calculated as a sum of partial products of commuters that use a specific transport type (public transport, cars, bikes, walk) and the pollution index assigned to the corresponding transport type; calculated by (12), where $C_p(i)$ is the corresponding pollution index, and the number of commuters is N_i, where $i \in \{pub, car, bike, walk\}$ presents the transport type

$$P = \frac{\sum\limits_{i \in \{pub,car,bike,walk\}} C_p(i) \cdot N_i}{N_{com}} \tag{12}$$

- *minimal noise pollution function NP*, calculated similar to the pollution function as a sum of partial products of commuters that use public transport, cars, bikes, or walk, and the corresponding noise pollution index; calculated by (13), similar to the pollution function, where $C_n(i)$ is the noise pollution index, N_i is the number of commuters in the transport type $i \in \{pub, car, bike, walk\}$

$$NP = \frac{\sum\limits_{i \in \{pub,car,bike,walk\}} C_n(i) \cdot N_i}{N_{com}} \tag{13}$$

- *maximum number of bikers*, a function that calculates the number of bikers, analysing the average numbers of commuters between any two determined city locations, and their preferences, which depend on the quality of the existing bike path, weather condition (such as sunny, bright, cloudy, rainy, windy) and temperature condition, mainly expressed by the average month temperature (for the Skopje case study the most bikers prefer April, May, September and October). The number of bikers is calculated by (14), where the capacity of a given bike path identified by i is $N_{max}(i)$, C_Q is the quality index of the bike path, C_m is the month average temperature index and C_w is the weather index depending on whether it is sunny, bright, rainy, cloudy or windy

$$N_i = N_{max}(i) C_Q C_m C_w \tag{14}$$

The constraints for finding the optimal solution are defined by:

- *values of all analysed indexes are within the range between 0 and 1*, calculated by (15), including the pollution index, the noise pollution index, the weather index, the month temperature index and the bike path quality index, meaning that the number of bikers can be between 0 and N_{max} for a given bike path that connects any two points in the analysed scenario, depending on user preferences and conditions that influent their transport.

$$C_Q, C_n, C_p, C_m, C_w \in [0, \ldots, 1] \tag{15}$$

- *the cost function of selected project proposals should fit in the budget*, calculated by (16), meaning that the sum of costs of selected project proposals is less than the allowed budget B, where K is analysed subset of project proposals P_j, where j is identification of the projects in K

$$\sum_{j \in K} C(P_j) \leq B, \tag{16}$$

A detailed elaboration of all minimisation functions, constraints, and definition of corresponding indexes can be found in [7].

This model can be calculated by exploring the set of all possibilities. However, this approach to find the optimal subset of project proposals is demanding a lot of processing and it is a time consuming function. Similar approaches might decrease the processing demands by using greedy algorithm alternatives.

V. Discussion about optimisation and the FUPOL approach

Besides the approach of solving the system of inequalities, the presented models may be analysed by a different approach. In this section we discuss the approach used by the FUPOL project based on simulation and visualisation.

The simulation used in the FUPOL is based on agent based simulation and the developed models are a good basis to apply this approach. The identified user groups which perform a different activity in the Vodno recreational model, or the commuters which are determined for their transportation demands in the Bike model can be presented by a corresponding agent. The agent can decide whether to use the recreational activities or use bike as transport means based on the weather conditions and outside temperature. In addition, the quality of the bike infrastructure and the existence of bike docking stations, or enabling intermodality may influence the user behaviour and decision, In the Vodno recreation model the user behaviour is also influenced by the conflicts that are modelled if two conflicting activities share the same resource. All these relatively complex situations are modelled and used in the agent based simulation.

The output of the agent based simulation is the number of users related to a given resource, or the resource occupancy determined in regular time slots. The simulation provides the output as a time series function per each resource.

The visualisation itself is using the output of the simulation. The average number of users per resource is presented on a graphical map and both the citizen and administration can simulate various scenarios, such as what happens if a certain project proposal is realised on the bike proposal, or if more users perform a certain activity on a given resource on the Vodno mountain. The visualisation enables citizens participate in the policy modelling by suggesting how to improve the schedule of recreation activities or increase the number of bike users. Feeling the effect to find out what happens if a given scenario is realised brings a virtual reality in a condition to improve the urban policy modelling.

Next we discuss the benefits and impacts of implementing the FUPOL project. Using tools from the FUPOL project can make efficient e-Government decisions for its future use. For example, FUPOL social media tools can collect public opinion and interest for various recreation activities and resources at given time interval. This will lead to an adjustment of the number and type of recreational activities. FUPOL simulation tools can support calculation of the optimisation functions, and also allow a visualisation of ideal, extreme, and optimal cases. This, in turn, FUPOL tools will bring ideas about resolving potential conflicts by defining appropriate schedule. Moreover, the simulation tool will be publicly available to anyone interested in seeing how Vodno is used. In this way, the citizens of Skopje can be educated on the impacts of their recreation. Also, they can be led to understand how and why certain decisions and laws for the recreation were imposed.

Recreational activities in a mountain close to a city increase the richness of services provided by urban ecosystems. The ecosystem services of a mountain such as Vodno could include the use of natural areas for recreation, and also amenity services such as aesthetics, that should be evaluated from four interdependent ”human dimensions” of greenways: cleanliness, natural preservation, aesthetics, as well as safety.

The planning of recreation activities inevitably involves trade-offs among the services provided to the different users and stakeholders. An efficient use of land should account for how the gain in one service compares to the loss of other services, and even whether the land is suited for a certain recreation activity or better left for development.

Next, we state the advantages of fostering bicycle intermodality in the city of Skopje. The benefits of the use of bicycles for flexible mobility are multiple, some of which include emission reductions, physical activity benefits, reduced congestion and fuel use, individual financial savings. However, there are some barriers that should be properly addressed to successfully promote the use of bicycle as an alternative to well consolidated transport means.

Gusev et al. present a more general approach about urban city modeling [8].

VI. Conclusion

Another approach based on simulation and visualisation is also used in the FUPOL project. This complementary approach uses the same input parameters and has the same overall goal, but is based on simulation of user behaviour instead of using predefined or calculated statistical values. Besides the fact that this approach will give a proof of concept, it will also enable an environment to involve citizens in the policy making processes.

Driven by the citizens demands, our research will lead to better policy decisions, more efficient implementation of government policies as well as better identification of consequences for citizens and businesses. With the help of multichannel social computing, policy topic sensing and extraction, advanced visualization including integration with GIS, multilingual semantic analysis, advanced policy modelling and model repository, dynamic agent based simulation, cloud computing and IMS supported crowd sourcing, we aim to address the bicycle intermodality problem, as well as the recreational activities scheduling at Vodno.

References

[1] J. D. Pintér, “Global optimization: software, test problems, and applications,” in *Handbook of global optimization*. Springer, 2002, pp. 515–569.

[2] I. Stefanoski, “Drafting the Bicycle Master Plan for Skopje,” City of Skopje, Tech. Rep., 2004.

[3] M. Gusev and B. Veselinovska and A. Guseva and B. Gjurovikj, “Future Policy Modeling: A case study – Optimization of recreational Activities at the Vodno Mountain,” in *Advanced ICT Integration for Governance and Policy Modeling*, P. Sonntagbauer, K. Nazemi, S. Sonntagbauer, G. Prister, and D. Burkhardt, Eds. IGI Global, 2014.

[4] IDORM, “Transport Master Plan for Greater Skopje,” City of Skopje, Tech. Rep., 2010.

[5] JP Ulici i Patista Skopje. (2014). [Online]. Available: http://www.uip.gov.mk/

[6] GUP, “General Urbanistic Plan of Skopje,” City of Skopje, Tech. Rep., 2012.

[7] A. Guseva, M. Gusev, and B. Veselinovska, “Fostering Bicycle Inter Modality in Skopje,” in *Advanced ICT Integration for Governance and Policy Modeling*, P. Sonntagbauer, K. Nazemi, S. Sonntagbauer, G. Prister, and D. Burkhardt, Eds. IGI Global, 2014.

[8] M. Gusev, G. Velkoski, A. Guseva, and S. Ristov, “Urban policy modelling: A generic approach,” in *ICT Innovations 2014, AISC 311*. Springer, 2015, pp. 187–196.

Recommending audio and video materials based on tag-based collaborative filtering

Aleksandar Kotevski, MSc
University St.Kliment Ohridski – Bitola
Faculty of Law
Bitola, R.Macedonia
aleksandar.kotevski@uklo.edu.mk

Cveta Martinovska Bande, PhD
University Goce Delcev - Stip
Faculty of Computer science
Stip, R.Macedonia
cveta.martinovska@ugd.edu.mk

Abstract— **The success of e-learning systems depends on selecting and providing adequate learning materials to learners, according to their requirements, need and goals. Additional, searching for adequate learning material in a large dataset without some techniques for filtering and recommendations is almost impossible and leads to inefficient learning process. In our previous research we have implemented an intelligent e-learning system that classifies students based on their learning style in order learning materials to be delivered in the most adequate format. According to VARK classification, students are categorized in 4 categories: Video, Audio, Read and Kinetics.**

In this paper we review the results of the implementation of a new module in our system for recommending materials based on tags posted from the students for video and audio learning materials. Students can post tags and timeframes for learning materials (video and audio) and receive list with suggested learning materials.

The focus is on students that belong to Video and Audio category based (based on VARK classification). The module uses item-based collaborative filtering and generates the recommendation list based on the student profile similarity and posted tags.

Keywords— collaborative filtering, tags, recommendation, learning materials

I. INTRODUCTION

The fact is that the computer technology is important part of all learning processes. At the same time the volume of available information and learning materials is rapidly increasing. This abundance of information has created the need to help students find, organize, and use resources that match their individual goals, interests, and current knowledge [1]. Personalized learning refers to e-learning systems that are developed according to educational experiences and works according to needs, goals, and interests of their learners. Ideally, recommender systems in e-learning environments should assist learners in finding relevant learning actions that perfectly match their profile, at the right time, in the right context, and in the right way, keep them motivated and enable them to complete their learning activities in an effective and efficient way [2].

It's common practice of using systems that are able to detect the learners' needs and find out the most adequate learning materials. At the same time it is very important because learning materials are available in different formats (text, audio, video, practical examples, external link, presentations and etc) on the one hand, and learners have different learning styles on the other hand. That's why, in [3] we have implemented an intelligent system for e-learning that suggests the most useful learning materials and deliver the learning material based on the most adequate learning style to the students. The system uses VARK questionnaire for determining the learning style and is incorporated in the educational process at the Faculty of Law in Bitola.

Based on VARK questionnaire, we are grouping students in four groups according to their perceptual preferences: Visual, Aural, Read and Kinesthetic. The focus of this paper is on the students from Visual and Aural category only. They receive learning material in video and audio format, respectively. The students can post tags, comments and time-stamps for all of the materials. In this research, we have implemented a module for recommending materials based on tags posted from the students for video and audio learning materials. The module uses item-based collaborative filtering and generates the recommendation list based on the student profile similarity and posted tags.

II. RELATED WORKS

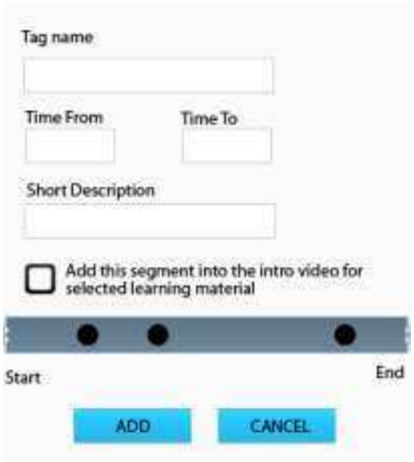
There are number of published papers that are focused on video materials in e-learning. The main objective in [4] is to highlight the importance and benefits of analytics and to support instructors with the appropriate resources for improving the use of their courses. The authors intend to combine and analyze learners' interactions with other available data obtained from learners, as such video analytics open new avenues for research on open and video-based courses. Authors in [5] propose e-learning system based on video tutorials and then present observations of how both groups of students interact with the multi video multimedia learning objects. Furthermore, they compare the results from students that used traditional course but with video tutoring and students that used distance learning course (watching and interacting more with the interactive multi video). In [6] authors propose a scheme for automatic video tagging, systematic representation and a proposal for a universal markup language for information exchange between devices including support for face recognition in videos on a distributed framework like social networking platform. In [7] authors describe a system for video sharing, commenting, and interest discovery that combines recommendation algorithms and tools for video tagging. In [8] authors focus on content-based tagging and user-tagged online videos. Also, they present an extensive benchmark using a database of real-world videos from the video portal youtube.com. They show that a combination of several visual features improves performance over their baseline system by about 30%. Authors in [9] describe DIVA, a decision-theoretic agent for recommending movies that contains a number of novel features. DIVA represents user preferences using pairwise comparisons among items, rather than numeric ratings, similarity measure based on the concept of the probability of conflict between two orderings of items. The purpose of the work presented in [10] was to develop a comprehensive, theory-based framework for creating instructional video podcasts designed to present worked examples. Sixteen design characteristics, organized according to four categories (establishing context, providing effective explanations, minimizing cognitive load, and engaging students), were used to develop 59 pre-calculus videos for 856 first-year university students. Overall, the vast majority of students noted that the video podcasts were useful and helped them understand mathematics better. Authors in [11] focus on the video learning research of the last years . In that content they take in consideration 166 related published academic papers. They sum up that the number of peer-reviewed articles dealing with video learning has been significantly increased during the last few years. Also research with mobile devices has significantly increased during the last years. In addition, the video learning research has moved to more asynchronous and non-interactive systems. In [12], authors make video analysis by providing a review of the current research in this area. The paper is focused on doing video-based reflection, using video cases to examine teaching practice, examine teaching practice in video clubs, using video editing to inform teaching practice, using video analysis (tools) to examine teaching practice and peer-video analysis as a form of reflection. Also it suggests that more research is needed to expand author's understanding of how evidence-based arguments are used to contribute to teachers' professional vision. Authors in [13] outline their initial investigations of applying information visualization techniques to lecture capture video systems. Their main goal was to better understand how students use these systems, and how visualizations can affect useful learning analytics. In that content, they applied three different methods to viewership data aimed at understanding student re-watching behavior, temporal patterns for a single course, and how usage can be compared between groups of students.

III. VIDEO TAGGING AND SYSTEM FUNCTIONALITY

Two important challenges for e-learning systems are the creation of targeted and personalized content for the users and selecting the most interesting and useful materials from the huge amount of learning content. In order to save student time for searching the most adequate video and audio content, we use two types of tags: audio-related and video related. Both contain few segments: title, timestamp and description. First type is intended for video materials only, while the other for audio materials available in the e-learning system implemented at the Faculty of law - Bitola.

Quickly accessing the contents of a video and audio is another challenge of this paper. That's why we have implemented user interface where students can specify tags' timestamp when they add or edit some tag. Then, when system will suggest learning materials to the students, it can redirect student at the exact timestamp. It is very user-friendly, especially in case of long audio or video materials because students will not waste time to find the exact time period in the suggested audio/video material.

Furthermore, there is an option for generating intro video/audio for materials, based on tags posted from the students. It means that while adding new tag, student can select the optional checkbox called "Add this segment into the intro video/audio for selected learning material". This is important for students while they review learning materials. Going through the intro video/audio, they will have information for the most important section of learning material.


Tag name

Time From

Time To

Short Description

□ Add this segment into the intro video for selected learning material

Fig. 1. User interface

IV. GENERATING RECOMMENDATION LIST

Collaborative tagging systems become very popular, practical and effective. In that context, tags could bring interesting and useful information to enhance RS algorithms [14]. In this paper we use three-dimensional relation: student – learning material (video/audio) – tag. According to the relation, the system knows which student has posted tag for specific learning material. We can define the following sets:

S = { S_1, S_2, ... S_n}: set of students

VA = {VA_1, VA_2, ... VA_n}: set of learning materials (video or audio)

T = {T_1, T_2, ... T_n}: set of tags posted from students S for learning materials VA

CheckTag (S_i, T_j, VA_k) = {false, true} is a function that requires three parameters: student, tag and learning material (video or audio). The function can return two possible values: true if the student S posted a tag T for the learning material VA or otherwise, can return false.

Let suppose that TSi is a set of tags posted from the student Si, where i = 1 .. n.
Then, $TS_i = \{t_j | t_j \in T, \exists VA_k \in VA, E(S_i,t_j, VA_k) = 1\}$, $Tu_i \subseteq T$, where VAS_i is a set of S_i, $VAS_i = \{VA_k | VA_k \in T, \exists\ t_j \in VA, CheckTag(S_i,t_j, VA_k) = 1\}$, $VAS_i \subseteq VA$, where TVA_i is a relation between tags posted from student and a set of learning materials.
$TM_i = \{<t_j, p_k >| t_j \in T, VA_k \in VA$, and $CheckTag (s_i,t_j,VA_k)=1\}$, $SF_i = (Ts_i, VAs_i, TVA_i)$ is defined as part of a student model S_i. General, user profiles can be defined as SF, $SF = \{SF_1 i=1..n \}$.
The system calculates the similarity between learning materials with the following function:
$Similarity(va_i,va_j) = wVAS * VASSimilarity(va_i, va_j) + wVAT * VATSimilarity(va_i, va_j) + wVAST * VASTSimilarity(va_i, va_j)$.
There, wVAC, wVAT, wVACT are with similar values. The sum of all similarities is equal to 1.
The similarities used in previous formula can be calculated as follows:
VATSimilarity(vai, vaj) is a similarity of the student tag-material relation and it's defined as a percentage of the common relations of two students:

$$VATSimilarity(va_i, va_j) = \frac{|TVAi \cap TVAj|}{\max\{TVAk|\}}$$

Where $TVA_к$ is a set of tags for the learning material VA_k while $Tva_k = \{t_i| t_j \in T, Similarity(va_k, t_j)=1\}$.

-VASSimilarity(vai, vaj) is a similarity of two learning materials. It's defined as a percentage of common tags from the same student.

$$VASSimilarity(va_i, va_j) = \frac{|SVAi \cap SVAj|}{\max\{SVAk|\}}$$

Where $SVA_к$ e is a set of learning materials that got tag VA_k from the student S, while $Sm_k = \{s_i|s_i \in S, \exists\ t_j \in T, Similarity(s_i,t_j, va_k) =1\}$
-$VASTSimilarity(va_i, va_j)$ is a similarity between two learning materials, and is calculated based on the percentage of the common tag-material relations:

$$VASTSimilarity(va_i, va_j) = \frac{|SVAi \cap SVAj|}{\max\{SVAk|\ \ \}}$$

Where SVA_j is a set of tags tj used by the student S, while
$SVA_j = \{<s_i, va_k>| s_i \in S, va_k \in VA$, and $Similarity(s_i,t_j,va_k)=1\}$

The list of recommended learning materials for the students can be generated after completing the following steps:

1) Determine the similar profiles to the logged student
2) Find out the similarity between learning materials used by selected profiles from the step 1 and select the top N learning materials with the highest level of conformity
3) The parameter N is actually a global variable manageable from the system administrator panel. The default value is 10
4) The level of conformity can be calculate with

$$SS_{va} (s_i, va_k) = \sum Similarity(va_k,va_k)$$

V. USE CASE

This section presents an example of how the system can be accessed and used by students:

- Step 1: Student logs in with his profile
- Step 2: The system checks the students' profile and gets information for his learning style, knowledge level and learning area
- Step 3: The system generates learning materials according to the step 2 and shows them in adequate format (video or audio) based on the most adequate students' learning style (step 2). As we mentioned before, all learning materials are cathegorized as video or audio
- Step 4: Student can go through introduction video/audio before watching/listening learning materials
- Step 5: If student finds useful information into the intro material, he can use full material and post tags by using tag interface
- Step 6: Based on posted tags from the student, and using tag-based collaborative filtering, the system detects all similar students, checks for materials they were using and generates a list with suggested learning materials
- Step 7: Student can select some of the suggested materials or next one from the list generated in step 3)

VI. RESULTS

The system was implemented at the Faculty of Law in Bitola. The obtained results are based on tests performed on three courses from undergraduate studies: Informatics, Constitutional Law and Introduction to EU Law. 247 students participated in the experiments. The students belong to categories Visual or Audio learner based on VARK classification. Particularly 110 of the students belong to Visual category and receive learning materials as videos, while the other students belong to Audio category and receive learning material as audio stream.

The implemented system contains 138 learning units in total, each of them composed from video, audio, text and examples and demonstrations. While using the system, the students from the video category set total 214 tags for learning materials for Informatics, 184 tags for Constitutional Law and 141 for Introduction to EU Law. Table 1 shows the results for:

- Number of learning units
- Number of tags
- Tags per student
- Tags per learning unit
- Number of students
- Used intro materials per student

- Used full materials per student
- Used learning materials from suggested list (%)

TABLE I. STUDENT ACTIVITIES

Activity	Category	Subjects		
		Informatics	*Constitutional Law*	*Introduction to EU Law*
Number of learning units	Video	88	26	24
	Audio	88	26	24
Number of tags	Video	214	184	141
	Audio	232	198	157
Tags per student	Video	7.64	4.18	3.71
	Audio	5.95	3.89	3.34
Tags per learning unit	Video	2.43	7.08	5.88
	Audio	2.64	7.61	6.54
Number of students	Video	28	44	38
	Audio	39	51	47
Used intro materials per student	Video	78.6	23.4	23.2
	Audio	74.4	24.4	23.8
Used full materials per student	Video	74.2	22.6	22.1
	Audio	72.8	22.8	21.9
Used learning materials from suggested list (%)	Video	76.4	71.18	74.43
	Audio	77.2	73.23	73.47
Used time-frames tags info (%)	Video	66.92	68.84	70.92
	Audio	71.2	70.47	74.24

VII. CONCLUSION

Important aspect in e-learning is system ability to select the most adequate learning content and deliver it in the adequate format to the users. The main goal of the system is to recommend the most appropriate materials to the students based on their tags for video and audio materials.

Another important aspect in this paper is video and audio introduction for each learning material, based on posted tags and timeframes from the student.

Based on the results, we can conclude that intro materials help the students, to decide what materials to use for learning. Because of that the number of used intro materials is higher than number of used full materials for both, video and audio.

Also, we can conclude that students were using time-frame information available for tags, so they skip audio/video materials to the exact time period.

According to the results of the examination of the students and answers given in the survey we can conclude that using collaborative filtering based video and audio related tags has positive effects and leads to more effective learning process.

REFERENCES

[1] R.Farzan and P.Brusilovsk, “Navigation Support in a Course Recommendation System”, 4th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, 2006, pp. 91-100

[2] T.Tang and G. McCalla, “Smart recommendation for an evolving e-learning system: architecture and experiment”. International Journal on e-Learning, 4(1), 2005, pp. 105 – 129.

[3] A.Kotevski, C.Martinovska Bande, R.Kotevska, “Learning style determination in e-learning system”, International conference of young scientists – Plovdiv, 2013

[4] N.Michai, G.Giannakos, K.Chorianopoulos, M.Ronchetti, Peter Szegedi and Stephanie D. Teasley, “Video-Based Learning and Open Online Courses”, International Journal of Emerging Technologies in Learning, Vol 9, No 1, 2014

[5] C.C. Viel, K. R. H. Rodrigues, E. L. Melo,R. Bueno, M. G. C. Pimentel and C.A.C. Teixeira, “Interaction with a Problem Solving Multi Video Lecture: Observing Students from Distance and Traditional Learning Courses”, International Journal of Emerging Technologies in Learning, Vol 9, No 1, 2014

[6] A.Sharma, B.Chheda, K.Karoth,T.Jain, ”Automatic video tagging system for a distributed framework”, CIS 6930 distributed multimedia systems, http://www.cise.ufl.edu/~ bnchheda/ automatic video tagging system for a distributed framework.pdf

[7] M.Bertini, A.Del Bimbo, A.Ferracani, F.Gelli, D.Maddaluno, D.Pezzatini, “A novel framework for collaborative videore commendation, interest discovery and friendship suggestion based on semantic profiling”, ACM International Conference on Multimedia (MM) - DEMO Session, 2013, pp. 451-452

[8] A.Ulges, C.Schulze, D.Keysers, T.M. Breuel, “Content-based Video Tagging for Online Video Portals”, http://www.keysers.net/daniel/files/youtube-muscle07.pdf

[9] H.Nguyen, P.Haddawy, The Decision-Theoretic Interactive Video Advisor, Conference Name Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999

[10] R. H. Kay, “Developing a Framework for Creating Effective Instructional Video Podcasts”, International Journal of Emerging Technologies in Learning, Vol 9, No 1, 2014

[11] N.Michail. G.Giannakos, L.Jaccheri and J.Krogstie, “Looking at MOOCs Rapid Growth Through the Lens of Video-Based Learning Research”, International Journal of Emerging Technologies in Learning, Vol 9, No 1, 2014

[12] M.Michael. P.Scott,”Digital records of practice: A literature review of video analysis in teacher practice”, Society for Information Technology & Teacher Education International Conference, 2012

[13] C.Brooks, C.Thompson,J.Greer, “Visualizing Lecture Capture Usage: A Learning Analytics Case Study”, Workshop on Analytics on Video-Based Learning (WAVe) , 3rd Conference on Learning Analytics and Knowledge, 2013

[14] H.Karen, L.Tso-Sutter, L.B.Marinho and L.Schmidt-Thieme,”Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms”, Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim, 2008

Social networks and social media in general – potentials for advertising and methods for measuring its effectiveness

Margarita Stoshikj, Ilijana Petrovska
School of Business Economics and Management
UACS
Skopje, Republic of Macedonia
m_stosic@hotmail.com, petrovska@uacs.edu.mk

Veno Pachovski
School of Computer Science and information technology
UACS
Skopje, Republic of Macedonia
pachovski@uacs.edu.mk

***Abstract*— The widespread and ubique presence of Internet and everyday usage of social networks and social media in general, results in a mind shift in advertising, as well. The marketers should go to where the population and audience are present, and the audience is becoming more present at the social networks. Therefore more and more companies are beginning taking advantages of new marketing opportunities of social media tools. In this paper researched is the level of using social networks as social media in companies marketing communications activities depending from the target group. The second important issue for companies to decide whether to allocate budget at social networks are the results from the campaign. The publicized research so far shows that different methods can be applied to different companies and there is no unified method that can be applicable for an arbitrary company. Also research shows that Macedonian companies do not measure financial aspect of social media ROI in general. This paper will be beneficial both for the scientifically research in defining the most appropriate method for measuring the Social media results and also for business in order to include measuring social networks as social media in their marketing communications campaigns.** (*Abstract)*

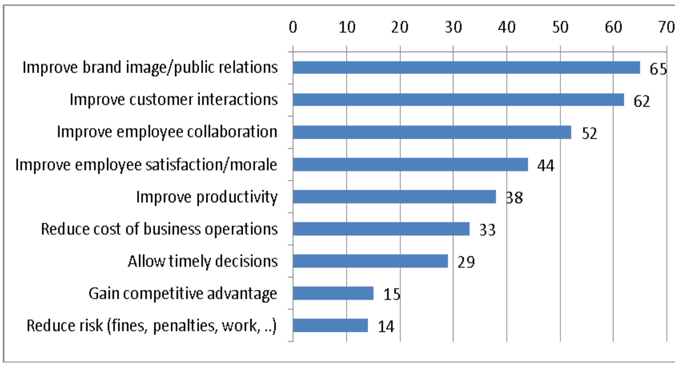
Keywords*— *social networks, social media, measuring social media effectiveness, ROI

I. INTRODUCTION

The marketing possibilities that Internet offers are growing each and every day, and as every process, it has evolution phases. The usage of some of the online marketing tools is in the decline and some are utilized more. Thus, the use of interruptive marketing, in a form of different formats, print, TV, direct mail, online ads, etc is in decline and considered obsolete.

Internet became a platform where one person can communicate with thousand others and very quickly, practically overnight, that communication started including services and products that companies can offer, thus becoming a platform that transformed marketing in general.

As mentioned in Chaffey et al, *(*2009) it became an opportunity or a threat to organizations. For customers, it offers wider choice of services and products, with differentiating prices from various suppliers from all around the world. For organizations, it provides an ideal opportunity to expand their market globally, apply newly introduced communication techniques and compete with large companies on equal footing. Following expansion of social media (Facebook, Twitter, LinkedIn) many companies rushed to explore the link between social media and their business.



Source: SocialTaps.com (2012), *Benefits of social media*

Fig. 1 Benefits of social media

According to (Rodrigez, et al. 2012), only the forward-thinking companies have implemented social media in their complete business. Even more the whole strategies of some companies are based on the social media advantages. Another research, Geho *et al,* (2012) who have surveyed 3342 marketers, 90% of them answered that social media is important for their business, 80% answered that social media generated more exposure to their business, 59% saw reduction in marketing costs.

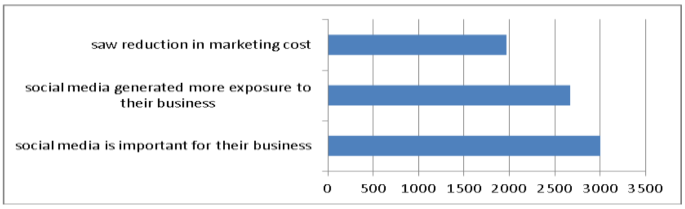
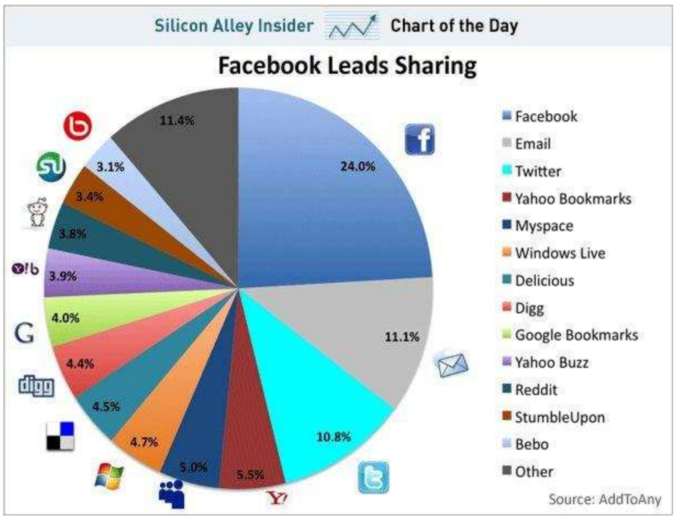


Figure developed by the author, based on Geho et al. (2012)

Fig. 2 Business gains

This research also shows that Facebook, Twitter and LinkedIn and blogs have dominating usage, and Facebook is a leader with 92% of usage.

Obviously, marketers have to be where their consumers are, and certainly these are social networks, states Shih, (2009). Social networks, sweeping across continents are a fast growing phenomenon and events over time confirm this finding.



Source: Ohhlson , J. (2011), Facebook Leads Sharing With 24% of Market

Fig. 3 Social networks Internet share

Social media supports the businesses, but to what extent that support is positive, or is it increasing effectiveness in social marketing program; still remains a subject of concern to many marketers, and needs to be further explored (Dickey and Lewis, 2010). Having that in mind, continuous search of marketers for an indicator of effectiveness of social media brought them to the concept of Social media Return on investment - ROI.

II. SOCIAL MEDIA ROI

Till 2010, many social marketers have claimed that social media could not be measured, and businesses should be satisfied with newly introduced interpersonal engagement, regardless of cost. Some of the past assumptions about social media still have influence at present, like those that social networking has success for fun and friendly products. Other assumptions are that benefits to the company are more on emotional level and that majority of users are teenagers and young adults.

The skepticism towards measuring social media ROI was stepping out in front of the need of more and more entrepreneurs to see financial benefit of money invested in certain social media campaign. Also, there is a need to measure how successful the company's marketing was in the past. This process can ease adoption of the budget and decision-making process inside the company.

For example, Powell *et al,* (2011) look at social media as "de rigueur" in near future, and in order to make right decisions, marketers need to measure effectiveness of social media in order to be certain that their marketing investments are spread optimally across all media, including the new trend of social media marketing.

And when a company knows the ROI for various social media activities, the marketing strategy can be easily refined, most of the problems diagnosed, budget allocations adjusted and the whole cycle repeated. Also, the measurement can help marketers share their results with others in organization in a form they understand and that is ROI.

A. Measurement of social media ROI

The need of managers to get real numbers for effectiveness of social media program has resulted in ready-made software and many methods offered by marketers for measurement of social media ROI. For example, ("Social media analytics", 2013) lists several most popular social media measurement tools nowadays:

- Hootsuite – ideal for managing social networks by tracking conversations and results of the campaign
- SocialBro- that can help every business grow through Twitter, analytics for social media presence and competitor tracking
- SproutSocial – measures Facebook impressions and twitter followers
- Google Analytics- gives social reports that measures impact of social media presence of the company
- SocialMention - is an analysis platform that "aggregates user generated content from across the universe into a single stream of information". It integrates results from Facebook, You Tube, Twitter, Google, Friendfeed in one place
- Gremln - it executes marketing campaigns on across different social platforms, and measures their performance
- SimplyMeasured – works with Facebook, Instagram, Google+, Tumblr, LinkedIn and many more. It measures "the cohesive relationship between each network, your competitors' social profiles, and how your audience interacts throughout the entire social space."

B. Financial aspect of measuring social media ROI

Some of the authors like (Ray. A, 2010) and Solis (2011), give credit to financial aspect of social media ROI, rather than qualitative aspect.

Although, some of the authors, like (Ramers, 2011), suggest that everything has to be measured and assigned real-money value, others suggest that only those parameters, which are aligned with the objectives of the companies have to be calculated and converted into money.

Due to the huge list of metrics (Solis B, 2011) states that the company has to choose those that are crucial for success of the program. In other words, the company has to establish KPI – Key performance indicator, and then decide which metrics can support its measurement.

The next important element is assessment of the costs. Calculating the cost of social media effort is a crucial phase before implementing measuring procedure. Costs can be divided as “hard” or “soft” costs. “Hard” costs are costs that we usually pay money for things like: software, service or product. “Soft” costs are difficult to determine, to mention just a few of them: salaries of people working on social media (hour rate, bonuses), time allocated for social media activities etc.

The financial data and ROI can be calculated with the following equation:

$$ROI=\frac{gain\_investement - cost\_investment}{cost\_investment}*100\% \qquad (1)$$

where

a) Gain_investment - the price for which a company sold the investment

b) Cost_investment - price that the company initially paid for the investment.

The same formula for measuring ROI is offered in Blanchard O, (2011).

It is said that ROI can be connected only with financial outcome and never with non-financial outcome because the unit of measurement (currency) must be the same in all parts of equation.

Another formula offered by Powell *et al,* (2011) give priorities to defining key drivers, (*to be measured for success*), that are considered to be in alignment with business objectives and benefits of the company.

In this case, the ROI equation which is usually used is

$$Mrktng\_ROI=\left(\frac{Inc\_Revenue*Contrib\_Margin\%}{Mrktng\_cost}-1\right)*100\% \qquad (2)$$

where

a) Inc_revenue – incremental revenue - a financial term that has several meanings. In its purest form, it simply is the increased revenue from a specified increase in sales (“What is incremental”, 2014)

b) Contrib_margin (CM)- Contribution margin - “the amount by which sales revenue exceeds variable costs (“Contribution margin, 2013”)

c) Mrktng_cost – marketing cost - “The total cost associated with delivering goods or services to customers. The marketing cost may include expenses associated with transferring title of goods to a customer, storing goods in warehouses pending delivery, promoting the goods or services being sold, or the distribution of the product to points of sale”. (“Marketing cost”, 2014)

Contribution margin can be calculated per unit or as a total contribution margin:

$$Unit\_CM = UnitPrice - VariableCostPerUnit \qquad (3)$$

where *variable costs are* “those which vary in proportion to the level of production”.

Variable cost may be direct as well as indirect. Direct variable cost includes direct material cost and direct labor cost. Indirect variable costs include certain variable overheads.

$$Total\_CM = TotalSales - TotalVariableCost \qquad (4)$$

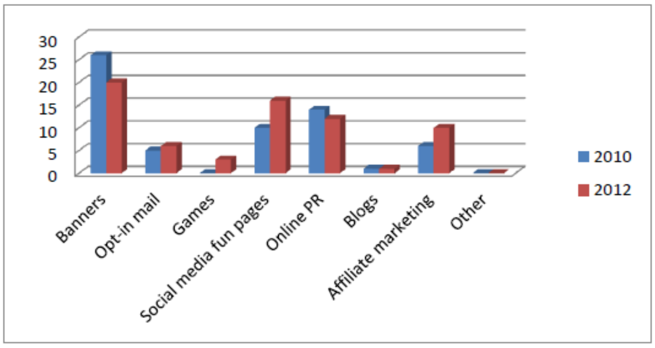
This measurement of ROI is needed when marketers usually present numbers outside the marketing team. But for internal use, only for marketing purposes they can use simpler equation:

$$ROMI=\frac{IncrementalRevenue}{MarketingCost} \qquad (5)$$

where Incremental Revenue and Marketing Cost have the same meaning as stated previously.

III. Situation in the Republic of Macedonia

The situation in Macedonia follows this trend, meaning that there is a shift within social media tools. For example in 2010 banners were most used by companies, and in 2012 there is a decline from 47% to 35%.



Source: [6]

Fig. 4 Social media used for advertisement

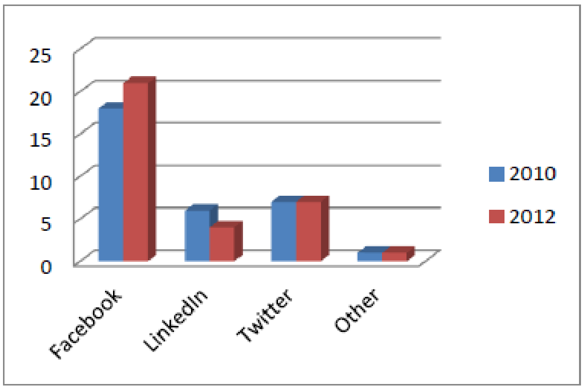


Fig. 5 Social networks in Macedonian companies

On the other hand there is general increase in using social media for advertising from 10% to 20%. It should also be noted that search engines register increase from 3% in 2010 to 7% in 2012 which is approximately 120% increase ("Trends in use of social media in Macedonia", 2012).

According to (Begu, B. 2011) the popularity of Social Media platforms as Facebook, Twitter is very high and increasing in Macedonia. However, it is mainly on entertainment level, opposed to the global trend, where customers expect feedback from companies for their engagement on social media.

IV. CONCLUSIONS

Social media and social networks are continuously changing, practically on a daily basis, and became integral part of everyday life and businesses in general. The traditional marketing gradually retreats in front of this new entrant (the social media), but still remains as a necessity and integrated part in the overall marketing strategy of the companies.

There is a variety of social media tools that companies can use for social marketing, but they need to select the most appropriate ones i.e. those are closely aligned with the company's goals.

Managers are willing to invest in social media activities, but would like to see the outcomes, converted to currency, of that kind of investment.

The research ("Trends in use of social media in Macedonia") shows that there is a shift in use of social media by companies in Macedonia.

Namely, there is an increase of 10% in the use of social media for advertising in the period between 2010 and 2012, but there is still no evident result that Macedonian companies measure social media ROI, as an indicator of success of its social media presence.

REFERENCES

[1] Blanchard, O "Menaging and measuring social media efforts". Pearson Education, Boston, 2011

[2] Begu, B "Social media and its impact on creating effective Marketing communications". University American College, Skopje, 2011

[3] Chaffey, D and Chadwick, F and Johnston, K and Mayer, R "Internet Marketing" pp.31 Pearson Education, Essex England, 2009

[4] Dickey, J and William F "The Evolution (Revolution) of Social Media and Social Networking as a Necessary Topic in the Marketing Curriculum: A Case for Integrating Social Media into Marketing Classes". Full Society for Marketing Advances Proceedings, p140-143, 4p 2010,

[5] Geho, P and Dangelo, J "The evolution of social media as a marketing tool for entrepreneurs", 2012

[6] Faculty of Economy, Skopje. "Trends in use of social media in Macedonia", 2012 [Accessed date: 15.01.2014] Avilable at: http://www.seebiz.net.mk/WBStorage/Files/Trendovi.pdf

[7] Powell, G and Groves, S and Dimos, J "ROI of social media". John Wiley & Sons (Asia), Singapore, 2011

[8] Ramers, J "Four Ways to Measure Your Social Media ROI", Corporate Meetings & Incentives, Dec 2011, Vol. 30 Issue 11, p26-26, 1p 2011

[9] Ray, A "The ROI Of Social Media Marketing". Forrester Research, Inc., 2010

[10] Solis, B "Social media measurement", 4imprint, Inc., 2011

[11] Social media TOP 10 analytic tools, Electronic version

Visualization and Learning of Macedonian Sign Language

*1Boban Joksimoski, *2Magdalena Kostoska,
*3Nevena Ackovska, *4Ana Madevska Bogdanova, *5Dragan Mihajlov

*Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
email: 1boban.joksimoski@finki.ukim.mk, 2magdalena.kostoska@finki.ukim.mk,
3nevena.ackovska@finki.ukim.mk, 4ana.madevska.bogdanova@finki.ukim.mk, 5dragan.mihajlov@finki.ukim.mk

***Abstract*—A significant number of people have inabilities with constructing and deciphering sound patterns, due to different medical conditions. We are focusing on the Macedonian sign language and advancing software tools for the deaf people in Macedonia. In Macedonia there are around 6000 deaf people and according to the National association of deaf and hard of hearing of Macedonia there are only 12 licensed interpreters of the Macedonian Sign Language (MSL).**

In this paper we present an idea of joint software platform dedicated to learning, visualization and interpretation of MSL. This platform will offer various methods of learning MSL, intended for all ages and all levels of hearing loss.

keywords - Macedonian Sign Language, Learning, Visualization

I. Introduction

According to the World Health Organization statistics updated in 2013, there are around 360 million people worldwide that have disabling hearing loss [1]. For those people, there is a significant challenge to integrate into the society. They represent a special cultural and linguistic group.

The method of communication for deaf and hard of hearing people is commonly known as Sign Language. As opposed to the common misconception, sign languages are related to natural languages and have wide varieties between each other. There is not a universal language and the languages differ from country to country. Sign language is also the first language of the people suffering hearing loss. Very often, these people have difficulties with written letters, because it's not their natural way of communication because written language is heavily influenced by auditory patterns. For them, the output should be presented in the form of gestures and expressions. The most acceptable way is by creating visual representations of the sign language they use. It can be accomplished using two techniques: video and/or image sequences and avatar based virtual signing.

Macedonian sign language is specific as it is not standardized, as opposed to other sign languages (for example American Sign Language), and its vocabulary is very limited (around 2800 signs including the letters of the alphabet) [2]. In Macedonia there are around 6000 deaf people and according to the National association of deaf and hard of hearing of Macedonia there are only 12 licensed interpreters of the Macedonian Sign Language (MSL). Unlike the bigger countries that promote their national sign language, there are no e-books, videos or any other type on online content that can help learning the MSL. Only one television channel (the national channel) offers once a day a MSL interpreted news and a weekly show called "The world of silence". These are the reasons why we started this project. We wanted to enter this world and to offer the possibility of introducing MSL to every citizen of our country.

The rest of the paper is organized as follows. Background and related work about current researches analysis of similar systems is presented in Section II. This research faces a lot of challenges described in Section III.One aspect of the platform is learning the SL through games and multimedia enhanced tools and this is presented in Section I. The second aspect, presented in Section V, is Content Management System (CMS) that offers dynamical management of the different image and video based sequences. The third aspect is avatar based visualization of MSL, exposed in Section VI. Each of these aspects is complementary to each other and together they create complete environment for MSL, which fully integration is explained in Section VII. Finally, in Section VIII we evaluate the significance of this platform.

II. Background and related work

Researches in this area include different points of view and activities: learning SL, signing SL and interpreting SL through moves recognition and lip reading recognition. Some researches focus on one area only, but sometimes they include two or more activities.

A big number of researches concentrate on the learning process of the sign languages. One research presents a HCI conceptual meta-environment framework to construct computational Intellectual Artifacts in SL to promote bilingualism (Libras-Brazilian SL/Portuguese) in order to increase family bonding activities and effective bilingualism for Deaf children and non-Deaf parents. This framework combines the SL with the cognitive theories (Sense Making, Common Sense, and Concept Maps). Its goal it is to give directions to design of

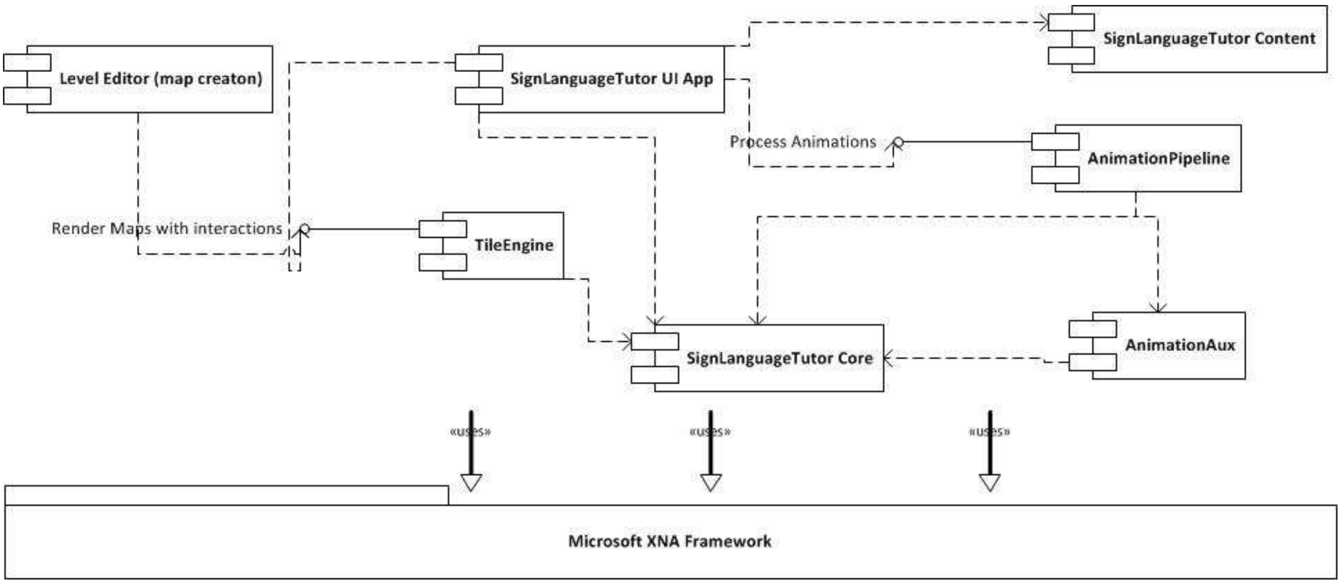


Fig. 1. Solution architecture

bilingual artifacts in various genres like games, educational objects, storytelling environments etc. [3]. According to another study of these authors, a pedagogical (teaching and learning) architecture should have strong pedagogical conception, a methodological systematization and should provide a virtual learning environment [4]. Another examples of this type of research is resulted in 3D Animation Editor and Display Sign Language System for Thai Sign Language [5], that exploits the XNA framework in order to train the general person and Deaf person with Thai Sign Language for ability to communicate to each other, 3D-based multi-model e-learning system for Chinese sign language [6] and SignOn, which is a Model for Teaching Written Language to Deaf People [7].

Some results of researches exploit the visual technologies for interpreting and improving the learning the sign language, like the System for Sign Language Tutoring [8] which evaluates users' signing and gives multimodal feedback to help improve signing. Other use robots, like NAO H25, to assist in teaching SL [9].

III. Challenges in creating sign language platform

Generating a general purpose and effective sign language platform is a massive undertaking. The main idea of all sign language platforms is creating all accessible tool that can narrow the gap between hard of hearing people and the rest of the society. A lot of research has been done in the subject of grammar of the sign languages (see [10] [11]), and it has been proven to be a difficult task. Furthermore, there are various ways of writing sign language, called notations, and each of them has its strengths and problems. The most used sign language notation systems are Stokoe[12], Hamburg Notation System[13] and Sign Writing. Most of these systems are adjusted for specific sign languages like the American Sign Language.

The Macedonian Sign Language poses a special challenge for visualization and interpretation. It has a valid grammar, but it is relatively poor in vocabulary (around 2800 terms, including letters and numbers). Furthermore, it is not standardized and, part of the work we are trying to achieve is to improve the standardization process.

During the creation of the desktop version of the application, one of the main challenges how to fully import the girl avatar, that was created in Autodesk Maya. We used Autodesk FBX XNA importer as a specific resource to import the 3D models, but we could not use the embedded resource processor in XNA because of the complexity of the 3D models (existence of animations based on skeleton movement attached to the model). That is why we have extended the embedded resource processor in XNA with AnimationProcessor module.

During the creation of the web version of the application we faced several challenges. The first challenge referred to the 3D model - we were not able to replicate similar animation rendering environment, so we were not able to use the animations in the original form. Additional challenge for the 3D model represents the size of the animations. The second challenge was the import of the games. We manage to find substitute web equivalent framework, but unfortunately it didn't support a big number of the original functions, so we had to rework big part of the games.

For the creation of the 3D based avatar visualization, it was necessary to rethink our whole approach to creating a web service that can effectively serve the clients. A substantial part of the work was creating and using different notations for all of 2800 present terms in the MSL.

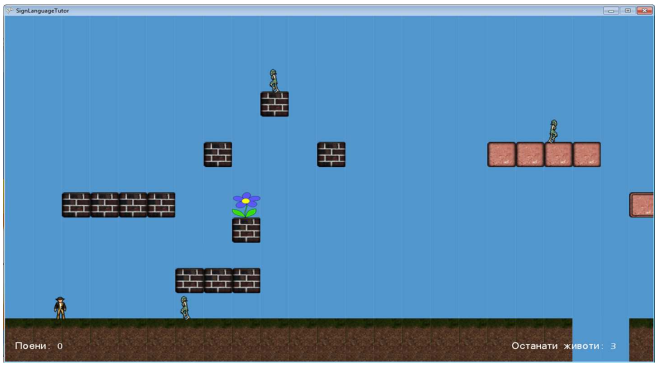


Fig. 2. Explore game - level 1

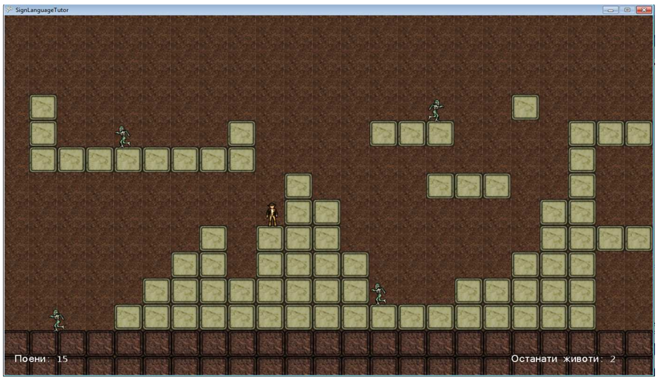


Fig. 3. Explore game - level 2

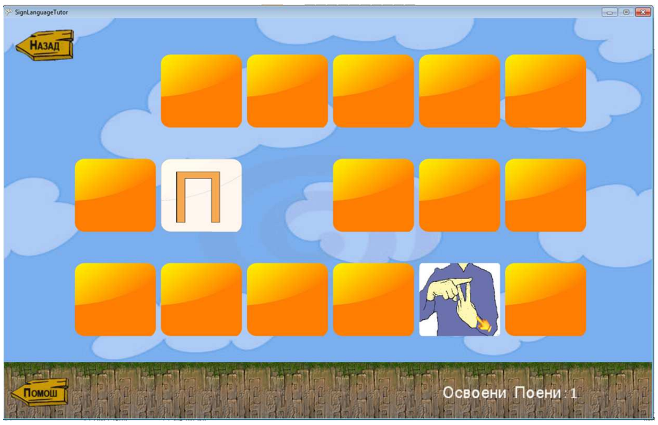


Fig. 4. Memory game

IV. Learning SL through games and multimedia applications

One of the most natural and easy-going approach for learning SL is through games. This is especially important for kids. We have created a natural environment for learning SL with the help of two-dimensional game named Explore and a memory game, using the Microsoft XNA framework.

Explore is a 2D game where our hero goes in an adventure against monsters and collects objects. Each of the collected objects is an award an animated 3D sign of that object in MSL. The game doesn't require many skills and it is adapted for kids. For these games two modules are created: TileEngine and LevelEditor, beside the main module and content, animation pipeline and animation auxiliary module. Fig. 1 depicts the architecture of the solution.

The SignLanguageTutor is the core module of Windows Game type and depends upon all the other modules. This module contains all the GameScreen Objects, as support screens as well as the core: MemoryMainScreen, TileMapScreen GestureSimulator. The TileEngine module serves as engine for customized independent game creation, more precisely creation of game map (tiles), movement options and view cameras. The map is created as two dimensional matrix, using three layers for each tile: front side, interactive side and back side. The LevelEditor module is XNA Windows Game application used for map creation, using the reference to TileEngine for real-time rendering and creates a serializable form of the game.

Figures 2 and 3 show the first level and the second level of the Explore module.

Memory is a module that represents a standard memory game where the goal is to merge an alphabet letter with the appropriate sign. At the begging of the game the MemoryMainScreen module randomly generated pairs of letter and signs. Fig. 4 shows the Memory module.

V. Content Management System for SL

In order to make the application available to a broader audience we have created a content management application for the signs of the MSL. The signs are divided into categories and subcategories and for each sign we have created gif files showing them from different perspectives in order to see the signing clearly. The system allows addition, modification and removal of signs, as well as categories.

As a base for sign creation we have used a girl avatar, created for the desktop version of the application. While in the desktop version we had an interactive environment to see every sign from the perspective the user chooses (shown in Fig. 5), we were unable to implement the same principle in the current version due to the limitations of the animation size and web engine. For that reason we have created gif files of every sign animation using two most appropriate perspectives, in order to best perceive the sign.

VI. Avatar based visualizations of SL

As a further enhancement of the multimedia applications, games and the sigh language content management system, the next logical step is creating a virtual three dimensional environment for visualization of the Macedonian Sign Language. The 3D environment extensively uses animation and game concepts for accurately generating sign language gestures and facial animations.

For the purpose of creating the virtual environment, we are using Three.js real-time 3D WebGL rendering engine to load three dimensional models to represent avatars. Currently we

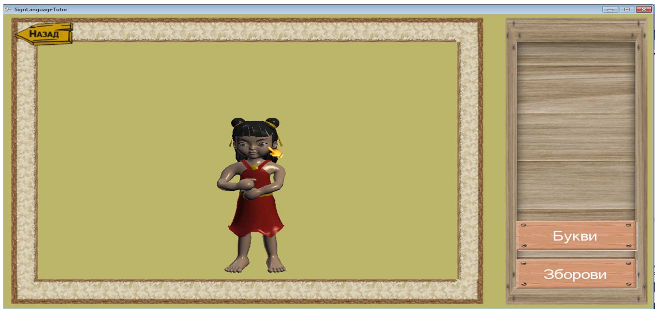


Fig. 5. Girl avatar in the desktop application

are focusing on adult male and female models, but accompanying children avatars are also in the work. Every avatar is endowed with expressive range of gestures and anatomical features as subtle as muscles for mimicking the changes in facial expression. For that purpose, the avatars represent anatomically correct humanoid 3D models, carefully modeled with muscle system in consideration. The avatar models are accompanied by a skeletal animation system, that is connected to the avatar mesh using smooth skin weight maps, that are capable of accurately replicating the upper body human anatomy.

The 3D rendering engine is communicating with a server-side Python web application stack and a PostgreSQL database stack, that is responsible for transforming the input data (currently text only), annotating it and converting it to skeletal control parameters that are passed to the client side and rendered using WebGL. Because of known issues present in 3D transformations, especially concerning rotations in 3D space, every transformation can be presented in various ways. For example, to escape the potential changing of axis directions, all of our multiple axis rotations are represented using quaternion parameters, instead of standard Euler notation. This method sacrifices human readability of the rotation parameters, but provides us with an gimbal lock escape mechanism. Part of the database schema, responsible for storing avatar transformations is given on Figure 6.

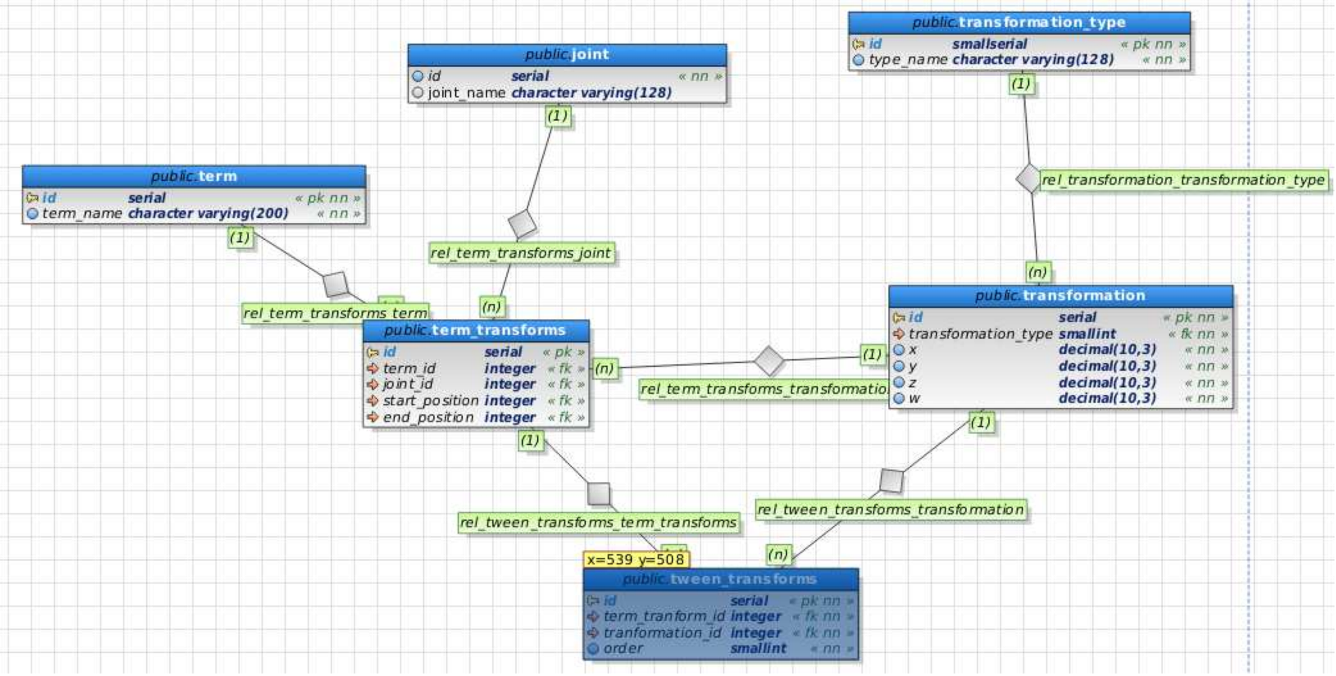
Furthermore, we are calculating the transforms of the joints for describing each individual gesture and we introduce extra positions for better interpolation between gestures. With the described architecture we are providing real-time sequences that can be "blended" together without creating noticeable flickering that is especially present with video or image base sequences.

VII. Future work

The integration process of the entire solution is massive. At this moment some of the modules are in mature stage and some are in developing or reworking process. We have defined several milestones for the future.

The first mile stone is integration of the MSL content management system and the created games. At this point this process demands working with several different frameworks and environments and it is in the testing phase.

The next milestone is integration and upgrade of the MSL content management system with avatar based simulation. This would add another environment to this already complicated system and would provide a fully integrated MSL platform.



public.transformation_type		
id	smallserial	« pk nn »
type_name	character varying(128)	« nn »

public.joint		
id	serial	« nn »
joint_name	character varying(128)	

public.term		
id	serial	« pk nn »
term_name	character varying(200)	« nn »

public.term_transforms		
id	serial	« pk »
term_id	integer	« fk »
joint_id	integer	« fk »
start_position	integer	« fk »
end_position	integer	« fk »

public.transformation		
id	serial	« pk nn »
transformation_type	smallint	« fk nn »
x	decimal(10,3)	« nn »
y	decimal(10,3)	« nn »
z	decimal(10,3)	« nn »
w	decimal(10,3)	« nn »

public.tween_transforms		
id	serial	« pk nn »
term_tranform_id	integer	« fk nn »
tranfomation_id	integer	« fk nn »
order	smallint	« nn »

(1)
rel_transformation_transformation_type
(n)
(1)
rel_term_transforms_joint
(n)
(1)
rel_term_transforms_term
(n)
rel_term_transfoms_transfomatio
(1)
(1)
rel_tween_transfoms_term_transfoms
(1)
rel_tween_transfoms_transformation
(n)
x=539 y=508

Fig. 6. Part of the PostgreSQL database schema responsible for storing the avatar transformation data

Furthermore, we are in the process of creating a web-based API that can be plugged to any Macedonian language website so that we can offer it as free service. This includes creating a text-to-sign translator that uses natural language processing techniques to infer the context and generate the meaning and convey it into MSL. Providing such interface can vastly improve the usage of Internet based services for the deaf and hard of hearing community in Macedonia.

VIII. Conclusion

Most of the develop countries offer easily accessible materials about native SL and their communities of deaf and hard of hearing people are open to the society. They constantly work on better inclusion. Unfortunately in Macedonia the situation is different. Information about MSL is poor and hardly accessible. The site of the National association of deaf and hard of hearing lack information and the deaf children are isolated in only two schools. One of the main aspect of this joint project is about raising awareness about the everyday needs and challenges that deaf and hard of hearing people and to offer bigger inclusion of this social group. Additionally this project follows the recommendations given by human-computer interaction research concerning deaf children about deaf children psychology and abilities

The possible uses of this kind of systems are vast. We are trying to accomplish couple of goals with the proposed system. The main objective is creating a digital database of all the gestures present in the Macedonian sign language, as a method to improve its standardization and accessibility. The second focus is to create a game learning system based on the Macedonian sign language and to promote the learning and integration of the Macedonian sign language among non-deaf people, so they can actively communicate with their children, deaf colleagues and friends. This is especially important for families and hearing parents with deaf children.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius" University.

References

[1] World Health Organization. (2014, February) Deafness and hearing loss. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs300/en/

[2] National Council of Disability Organizations of Macedonia, "Macedonian sign language dictionary," 2000.

[3] C. Guimaraes, D. Antunes, L. Garcia, A. Guedes, and S. Fernandes, "Conceptual meta-environment for deaf children literacy challenge: How to design effective artifacts for bilingualism construction," in *Research Challenges in Information Science (RCIS), 2012 Sixth International Conference on*, 2012, pp. 1–12.

[4] C. Guimaraes, D. Antunes, L. Garcia, L. Peres, and S. Fernandes, "Pedagogical architecture – internet artifacts for bilingualism of the deaf (sign language/portuguese)," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, 2013, pp. 40–49.

[5] P. Ittisarn and N. Toadithep, "3D Animation Editor and Display Sign Language System case study: Thai Sign Language," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 4, 2010, pp. 633–637.

[6] J. Liu, Y. Chen, Q. Yan, and J. Liu, "A direct3d-based multi-model e-learning system for chinese sign language," in *Digital Image Processing, 2009 International Conference on*, March 2009, pp. 100–104.

[7] M. Hilzensauer and F. Dotter, ""SignOn", a model for teaching written language to deaf people," in *IST-Africa Conference Proceedings, 2011*, 2011, pp. 1–8.

[8] O. Aran, I. Ari, L. Akarun, B. Sankur, A. Benoit, A. Caplier, P. Campr, A. Carrillo, and F.-X. Fanard, "Signtutor: An interactive system for sign language tutoring," *MultiMedia, IEEE*, vol. 16, no. 1, pp. 81–93, 2009.

[9] H. Kose, R. Yorganci, and I. Itauma, "Humanoid robot assisted interactive sign language tutoring game," in *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, Dec 2011, pp. 2247–2248.

[10] R. Elliott, J. R. W. Glauert, J. R. Kennaway, I. Marshall, and E. Safar, "Linguistic modelling and language-processing technologies for avatar-based sign language presentation," *Univers. Access Inf. Soc.*, vol. 6, no. 4, pp. 375–391, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1007/s10209-007-0102-z

[11] E. Efthimiou, S.-E. Fotinea, C. Vogler, T. Hanke, J. R. W. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and J. Segouat, "Sign language recognition, generation, and modelling: A research effort with applications in deaf communication." in *HCI (5)*, ser. Lecture Notes in Computer Science, C. Stephanidis, Ed., vol. 5614. Springer, 2009, pp. 21–30. [Online]. Available: http://dblp.uni-trier.de/db/conf/hci/hci2009-5.html#EfthimiouFVHGBBCMS09

[12] W. C. Stokoe, "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf," *Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3+. [Online]. Available: http://dx.doi.org/10.1093/deafed/eni001

[13] T. Hanke, "Hamnosys-representing sign language data in language resources and language processing contexts," in *LREC*, 2004, pp. 1–6.

Testing as a Service

Petar Vuksanovic*, Anastas Mishev
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
Email: petar.vuksanovic@gmail.com, anastas.mishev@finki.ukim.mk

Abstract. In the neverending search for the products with best quality, this study reviews some new ways of testing software systems and provides a proposal by seting the standards and guidlines that can improve testing effectiveness, time consumption and reduse costs. The focus is on the Testing as a Service principle which is an outsourcing model for testing that is capable of fast configuration of the test environment, high scalability and increased objectivity of the test results. The high level of expertise that comes from the fact that Testing as a Service providers are specialised in this field and have experience in designing many types of test environments brings added value to the subscribers of these services. In this paper first we will analyze existing Testing as a Service providers and we will provide a breef overview of all the possitive and negative efferct of this approach. In ordrer to create the proposal we will make an isight in one of most wide spread standards of the world – ISO 9001 and especially 9126 which is focused on software quality standards. The combination of this insight and the experience from the existing ogranizations and testing institutions brings us to the final results which is a set of standards and guidelines that every Testing as a Service provider should follow.

Keywords - Testing as a Service provider; testing effectivness; software systems; software quality standards.

I. Introduction

Today, applications are usually multilayered and are based on different platforms, technologies and programming languages. There are software systems that are made of thousands of lines of code which include very complex scenarios. [1] To make sure that everything works as it should there are several types of testing and analyses that should be made. These analyses can be very time consuming and expensive – things that most of the organization or individuals does not have.

On the other side, in this fast growing world of internet, customers want the products with most features and want them as soon as possible. Because of this, often, companies are releasing fast updates with code that hasn't been tested appropriately. The issues that will appear on software execution will surely have bad impact on company's brands. [2]

Testing as a Service principle has proved to be very effective in these kinds of situation - when the time for settings the test environment is short and high level of expertise and experience is needed so that all of the use cases of the software are covered.

There are many types of testing techniques but basically they can be sorted in these two types: black box testing and white box testing. White box testing is a detailed investigation in the internal structure of the source code. In order to perform this type of testing, the tester needs to have knowledge of the internal logic of the source code. Since this source code in many cases is protected, this technique of testing does not meet our needs. Black box testing is a technique where no knowledge of the internal implementation is needed. [3] When performing black box testing, testers will interact with the users or web services interfaces so this technique of testing is excellent for Testing as a Service providers.

Testing as a Service is a model of testing where software is tested as a service provided to customers across the Internet. This means that the subscribers to these serviced don't need to run tests on their computers which reduce the need for installing and maintaining the test environments. [4] Another benefit from this model of testing is increased objectivity of test results since the testers don't work together with the developers and they don't depend of the software system that is tested.

When subscribed to one of these Testing as a Service providers, companies are reducing costs of the testing since they pay for the package that suit their needs. These packages are sets of test that can be customized by the customers. The increased efficiency as a result of the experience of the providers in configuration of many types of testing environments, high scalability and the speed in which test environments can be installed brings added value to the customers.

This paper is organized as follows: in the second section an overview of few existing Testing as a Service providers is made and some of their attributes are listed. In the third section ISO standards are explained together with the reasons why organizations are certifying. In the fourth part the procedures, standards and guidelines that every Testing as a Service center should follow are explained. These procedures are organized in four steps: preparation, planning, test execution and analyses and include three types of testing: functional, load and performance testing. In the conclusion section a brief overview of the reasons why Test as a Service provides need to acquire this kind of certificate is provided

*) Graduate student at FCSE at the time of writing.

and how these set of procedure can improve the software development processes in the future.

II. Existing Testing as a Service providers

A. Oracle Testing as a Service

This solution is a cloud based platform and is used for delivery of automated testing services. It is designed for private clouds where it controls the testing process. The software provides rich application monitoring and data analysis. There are special charging facilities that are metering the usage and generate cost base on that usage. Some of the key featured are:

- Portal for executing load and functional tests
- Library that stores all test assets needed for the testing process
- Integrated application monitoring and diagnostics center[5]

B. HP Testing as a Service

This service is able to test software solutions in wide range from simple once to most sophisticated ecommerce applications. It supports many kinds of testing but is specialized on mobile applications. Some key features of this service are:

- Functional, usability, performance, security and localization testing for mobile applications
- Device testing where application is tested with help of different emulators so that customers can be sure that the application will work as desired on different devices
- Service testing which test the integration with services from telecom operators [6]

C. Winpro Testing as a Service

This service offers test packages to its subscribers with pay per use and outcome driven model. The goal of each testing is to make sure that the product is ready for the market. The packages are predefined using the experience that this service provider has in this field. Winpro services are separated in three fields:

- Certification, which guaranties that product is confirmed by a specific standard
- Testing service packs, which are predefined sets of tests by industry and specific domain
- Test environment and lab services, which are used to eliminate the CAPEX requirements. [7]

III. ISO standards

Procedure is a fixed, step by step sequence of activities with strictly defined start and end points that needs to be followed in given order to perform a task. [8]

Standards are documents that provide specifications and guidelines that products or processes need to follow in order to fit their purpose. [9]

A. ISO 9001

ISO standards are implemented in more that million counties in the world. Most wide spread standard of all ISO standards is ISO 9001 which is focused on quality management. Getting this certificate does not mean that companies need to make changes in the way they are managed.

B. Acquiring certificate

If the organization is famous in the field in which it operates, than the certificate won't add value to it. In opposite case, when the organization wants to expand in other domains and the customers does not have information for the ability whether the organization can respond to the challenge, than the certificate can add substantial value to it. [10]

C. ISO/IEC 9126

This is currently one of the best standards in the world for software quality. It is focused on quality on both software development and evaluation of software. This standard observes the quality by these four perspectives: quality model, external metrics, internal metrics and quality in use metrics. Here, quality model describes the relationship between different definitions of quality, internal metrics are those metrics that do not depend on software execution, external metrics are the once that do depend on software execution and are measuring the characteristics from the quality model and quality in use metrics which are metrics that are measured only when final product is ready and it is running in live environment. [11]

IV. Establishing Test as a Service center

For acquiring a Test as a Service certificate, the organizations will have to prove that they follow the set of procedures and standards that are defined for each type of testing. Because of this, every test procedure will have to be carefully explained in the final reports and these reports will be used as a proof in the revising process. Although these procedures are unique for each type of testing, they all are written in template which has four phases: preparation, planning, test execution and results analyses/reports writing.

Preparation is first of the four steps. These are the subtasks for preparation phase:

- Documenting the test procedure. This means that all the testers and engineers will have the documentation before the test process has started.

- Defining the goals of the testing which needs to explain the reasons for which this testing is needed and what the expected results are.

- Define the key metrics. There are several types of metrics that needs to be defined in one testing process. Test coverage is a metric which is used for measuring the test process progress and planning. Defect density is a metric that represents the number of defects found by product feature. If it is very high at some point, that than the cause of the density should be investigated. [14] When chosen carefully, these metrics will provide accurate information for the performance of the application. Also, this information can be used to identify potential bottlenecks, bugs and problems in that application. Finally these metrics are used to gather data which is then processed and compared with acceptance criteria.

- Defining the software tools and frameworks that will be needed for the test execution.

- Defining the resources. There are few types of resources. The first and most important resource are the employees that will be included in the testing process. Here we need to define the roles of every person, its obligations and responsibilities. Hardware is another type of resource that needs to be predefined; for example: required numbers of working stations, processors, RAM memory, hard drive memory, network bandwidth etc.

Planning is the second step. These are the subtasks of the planning process:

- Defining the test strategy. Everything related to possible risk, test schedule, test priorities have to be elaborated here. Another thing that needs to be defined is the communication routes. This means that all the communication routes between the management, testers and engineers needs to be predefined and that includes regular communication and cases when problems in test process will occur. [12]

- Defining the coverage criteria; this means that there will be predefined number of test statement that needs to be executed. This will also help in tracking the test process progress.

- Define exit criteria. A list of preconditions that need to happen so that test team can announce that testing process has finished has to be defined here.

- Defining the acceptable solution on how the whole test process will be monitored.

Test execution is the third step. The quality of all the previous steps can be diminished if this step is not executed properly. The simulation must reflect the test design so that the data that is collected from the test process are useful. When this is not the case, the results are error prone. This is a short list of guidelines that needs to be followed in order to maximize the test execution effectiveness:

- The test environment must be configured so that it is as similar as possible to the production environment. When the results will be analyzed, all the difference between the test and production environments must be taken into consideration.

- Only the tools that were previously listed in the preparation step must be used.

- Work stations on which the tests are executed must not be overloaded. It is also important that no other processes like antivirus applications, scripts or other applications are running on these stations so that they won’t have influence on the test results.

Analyzing the results and reports generation is the fourth and final phase. The managers and stakeholders will need lot more than just lists of test results. They will need conclusions. Analyzing the results helps to indentify the bottlenecks that might occur in high traffic. These are the subtasks of this process:

- Analyzing the data gathered from the testing process. This data is compared with the acceptance criteria of all the predefined metrics to make sure that the behavior of the application was satisfactory.

- Analyzing the metrics to find potential problems and issues in the application. If these metrics prove to be insufficient, than new test iteration should be considered where the list or metrics will be enriched.

- When these types of analyses are conducted it is common for huge volume of data to be gathered. Often, this volume is lowered so that the analyses can be made faster. This is a moment when lots of useful information can be lost so it is very important that the techniques for cutting the amount of data that is analyzed are carefully selected.

One reason why reports are made is the need for written evidence that the procedure was followed and that evidence can be presented to the revision commission if needed. There are several types of reports that can be made after the analyses are finished.

First type is technical report. This report contains description of all the tests and test environments. The data needs to be clear and it can contain explanations on advanced technical level. Access to all the test data and test cases needs to be provided. Tables are especially useful for this type of reports. They can provide a great deal of accurate information if they are effectively presented. Graphs are another way of providing accurate technical information. If graphs are chosen, they need to be clearly labeled with carefully identified scales. [13] In the end, short statements technical level concerns, questions and requests for collaboration must be provided.

Stakeholder report is the second type of report. The data which is elaborated needs to be clear and concise. It is especially useful if there is visual presentation of the acceptance criteria versus results gathered from the test process. Locations of all the documents, data and test cases are one component that this kind of reports must have. In the end a set of conclusions, recommendations and reasons for concerns are needed to finalize this report.

The key to effective reports is the presentation of the results to be fast, intuitive and interesting. Here are several steps that will make a difference:

- Reports should be sent often and sent to all participants in the test process
- Reports should contain intuitive visual representation of data.
- Reports should be customized for the intended audience.
- Reports should filter out the unnecessary data.

A. Functional testing procedure

This type of testing should confirm that all the functionalities that are listed in the requirement specifications are met. This is a short overview of the procedure for functional testing that every certified Test as a Service center must follow:

- The preparation for functional testing starts with test forms creation. The test process will run as planned only if the test procedure is predefined and explained concisely in well formatted documents. A copy of every form should be distributed to all persons that are included in the testing process. Every form must include the goal for the test and a short explanation with the reasons why the test is created and what are the expected results. This will ease the work of the employees that will conduct the testing.

- Instructions for how the testing process should be documented need to be provided. Here is should be explained how test results are documented when both test finished successfully and problems occurred.

- The roles and responsibilities of every person that is included in the test process needs to be predefined. These roles depend on the type of software that is tested.

The next phase is planning. The first step of this phase is writing the requirements in each test form. This will help in understanding the reasons why the particular test should be conducted. Another important step here is defining the preconditions. They are very important so that the testers start the test process only when everything is ready. The list of preconditions has two components – first one is a big picture on the global preconditions for the test process and the second one is set of preconditions specific for each specific test. Explaining the procedure comes afterwards. A step by step explanation is needed. The level of details depends of the requirements.

Conducting the functional test is the third phase. At this point, all the forms need to be completely filled with data. Every deviation from the test procedure must be noted. After every test execution, the initials of the person who run the test must be written so later, if retesting is needed at some point, that same person is contacted for opinion recommendations. After the testing process, a complete set of test forms must be kept as a valuable reference and it can be used as an evidence for revising process.

A final report for functional testing is the last phase. Every action that occurred in the testing process needs to be recorded. Also, a list of all the problems that occurred in the test process needs to be provided and explained. If some bypasses of the procedure were a must, they too must be noted in these reports.

B. Load testing procedure

Load testing is used to identify the maximum overload that one application can stand while its response is satisfactory compared to the acceptance criteria metrics. These are the key tasks for the preparation phase.

- Identify the load test key metrics. There is huge number of metrics that can be collected in one test process. Gathering information for all these metrics can have a really bad impact on the test results. That's why it is very important to identify the key metrics that will add most value to the process itself. Defining a set of questions that are related to the application performance and then using metrics to answer those questions, has proved to be a very effective approach.

- Identify the load test key scenarios. These scenarios represent expected users activities. Key scenarios are the once that require significant amount of primary resources and thus increase the risk for undesired application performance. These primary resources usually include: processor, hard disc memory, RAM memory and network bandwidth. Log files can be used in detecting the most used scenarios.

- Identify the test design. Tests designs can be made only after key metrics are selected and load levels are specified. Every test must have a specific purpose. The final goal of every test should be collecting enough information that will allow test team to analyze, configure and evaluate the software application.

In the phase planning the load testing process, these are the key steps:

- Identify the load test acceptance criteria. These are the parameters which are especially important here: response time, number of actions, resource utilization, maximum user load and business related metrics.

- Identify the load levels that will be used in the test scenarios. In this step a prediction is made for the load in the production environment. That way the test results will be more reliable.

Functional test execution must reflect the test design so that simulation is usefull. In other case, the test results are not reliable and thus, not valuble. The test environments should be configured so that is is as similar as it can be to the production environment. The tools that will be used for load generation must be the once chosen in the preparation phase. In the start of the simulation the application should be observed under minimal load and the load amount should be slowly increased until it reaches the final level that was defined in the planning phase. The load should be increased slowly so that system will has time to adapt.

Results analyses can help in finding bottlenecks of the software performance. These analyses require training and experience with dealing with big amounts of data so that valuable conclusions can be made. The analyses are conducted by comparing the collected data with the acceptance criteria and the metrics that were defined in the planning phase. All the steps and conclusions made from these analyses need to be clearly explained in the final report.

C. Performance testing procedure

Performance testing is a type of testing where functionality of whole system and its components is tested in a given set of conditions. Parameters like resource consumption, scalability and robustness can all be measured in this process. These are the key steps in the preparation phase for performance testing:

- Identify the key performance metrics. The data that is collected through these metrics will be compared with the desired performance characteristics in the final phase. When defining these metrics, the fact that external factors like users, network bandwidth and other systems will influence the metrics must be taken into consideration.

- Identify the key performance scenarios. The key scenarios in performance testing are usually derivates from the process of defining the acceptance criteria. If this is not the case, than these scenarios needs to be defined explicitly. There are several types of key scenarios: contractually obligated usage scenarios, key business scenarios, most frequent scenarios, performance intensive scenarios etc. [15]

- Configuring and designing the test environment. One important thing to consider is the different types of users that will use the software and their deviations from the use cases that were predefined. The final goal is minimizing the differences between the test and production environment.

In the planning phase, these are the steps that must be followed:

- Identify the performance test environment. This includes identifying the production environment, physical test environment and testing tools. The physical environment includes hardware, software and network configuration.
- Identify the acceptance criteria and how they can be used for performance evaluation. They can be also used to find the combinations of configuration parameters that will maximize the performance characteristics of the software.
- Identify the test tools and techniques that will be used to monitor the whole testing process.

The third step is test execution. This process must be coordinated and carefully monitored by the team members. While running the tests, validation of the configuration and test data is required in short intervals. After the tests are finished, a quick overview of the collected data is needed for some obvious deviations. All the test data and tests needs to be archived because sometimes retest must be conducted.

Through results analyses, collected data will be compared with the acceptance criteria of the performance testing. Sometimes these analyses can show that the performance level is not satisfied. In this case, modifications and optimization of the software are needed. After these modifications retesting is a must and in the retesting phase, a check for new cases that may arise as a result of the modifications should be conducted. Every step of the results analyses needs to be clearly explained in the final reports.

I. Conclusion

This paper outlines the benefits that one organization can have from subscribing to Testing as a Service provider; that is when testing is conducted by outsourcing organization, companies will be able to stay focused on their products and will also benefit from decreased testing time and lowered cost. Additional benefit from this approach is the fact that these Testing as a Service providers have experience in many types of test environments, can offer high scalability, and will increase the objectivity of the test results.

The set of procedures and guidelines in a form of certificate will maximize the effectiveness of the providers and will also decrease the risk of the subscribers to these services. These procedures were written by taking the experience from the experts in the field of testing software solutions and most popular software quality certificates and standards which in other words is a combination of practice and theory which has proved to be successful in huge number of cases.

References

[1] How to Deliver Resilient, Secure, Efficient, and Easily Changed IT Systems in Line with CISQ Recommendations". Retrieved 2013-10-18.

[2] Importance of Testing, June 1st 2012 http://www.softwaretestingclass.com/importance-of-testing/

[3] Penetration testing for web services - IEEE Computer society magazine, Feburary 2014; Nuno Antutes, Marco Vieira; pp. 30-31

[4] Software Testing as a Service (STaaS), Leo van der Aalst, Solution and Innovation manager at Sogeti Netherlands B.V. http://www.tmap.net/sites/tmap.net/files/attachments/Paper_STaaS.pdf

[5] Oracle Testing as a Service (TaaS), Oracle official site, 2013 http://www.oracle.com/technetwork/oem/cloud-mgmt/testing-as-a-service--1905801.html

[6] HP Software Professional Services for mobile testing , May 2012 http://h20195.www2.hp.com/V2/GetPDF.aspx%2F4AA4-1014ENW.pdf

[7] Wnpro official site, 2013, http://www.taas.wipro.com/index.aspx

[8] Business Dictionary, 2013 http://www.businessdictionary.com/definition/procedure.html

[9] ISO official site – „Standards“, 2013 http://www.iso.org/iso/home/standards.htm

[10] ISO 9000 Quality System Handbook, Sixth Edition, David Hoyle, 2009

[11] ISO Software Quality Standards and Certification, Anastas Mishev, Bisera Dugalic, Ss. Cyril and Methodius University - Skopje, 2012

[12] Communication in Testing: Improvements for Testing

Management, Tuula Paakkonen and Jorma Sajaniemi, Department of Computer Science and Statistics, University of Joensuu

[13] A guide to Technical Report Writing, IET The Intitution of Engineering and Technology, 2012 http://www.unlv.edu/sites/default/files/24/ Engineering- GuideTechnicalWriting.pdf

[14] Useful Automated Software Testing Metrics, Thom Garrett, IDT, LLC, September 2012 http://idtus.com/wp-content/uploads/2012/09/ UsefulAutomatedTestingMetrics.pdf

[15] Web Application Performance Testing Core Activities, Performance Testing Guidance for Web Applications, J.D. Meier, Carlos Farre, Prashant Bansode, Scott Barber, and Dennis Rea Microsoft Corporation, September 2007 http://msdn.microsoft.com/en-us/library/bb924359.aspx

Session 7

Intelligent Systems, Robotics, Bioinformatics 1

GIS classification for water quality estimation

Andreja Naumoski, Kosta Mitreski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
andreja.naumoski@finki.ukim.mk, kosta.mitreski@finki.ukim.mk

***Abstract* — Using the machine learning algorithms, it is possible to extract useful knowledge from ecological data. This information, can be not only used for explaining the ecological system, but also can be used to reveal the cause of the specific processes. In this direction, the paper focuses on using the fuzzy classification algorithm, namely the pattern trees to obtain new knowledge from the diatom measurement dataset. The fuzzy diatom models uses advantages of the fuzzy theory and ability of the predictive modelling to predict the outcome of given condition for given diatom. Fuzzy modelling depends from many factors, mainly by the shape of the membership functions, similarity metrics and the operations between the fuzzy sets – the fuzzy operators. In this paper we introduce new fuzzy operator (fuzzy geometric operator) that considers another property of the ecological datasets – the ration measurements. Important factor in the process of learning the model plays the first part of the algorithm that transforms the input set from crisp values into fuzzy values, and then continues the induction of the tree. The transformation is achieved by using different membership functions, which have different shape and mathematical description. This is very important, because later in the induction phase this will have effect on the classification accuracy and complexity of the obtained model. So, using this metric property, we will not only represent the model by the algorithm itself, but also it will be represented in spatial model using some software for geographical representation. The evaluation results are presented in the paper to compare the fuzzy operators and the membership functions with different evaluation criteria.**

Keywords—Classification models, GIS, membership function, water qaulity, modelling

I. INTRODUCTION

The ecological data contains very useful information about the processes inside of the ecosystem that can be used to describe and analyse the system. The extracting process is also very important, in manner of quality of information that is measured, but also the quality of the methods used to processes the measured data. It is important to note that the measurement process should be done in collaboration between the ecological scientist and the computer engineer in order to get high quality data. After the measurement is done, there is little that can be done to improve the knowledge extraction except to pre-process the data and try to remove noise from the measurements. The use of these analysis with the proposed methods can be used in many area of research, especially the diatoms with the similar diatom composition or physical-chemical parameters [1]. These methods can be used in several ways, some of the usage of these methods is to detect groups of samples with similar diatom composition or with similar environmental features. So we can use this to detect indicator species for the groups or relating biologically based groups to environmental variables. These variables that characterise the ecosystem in terms of quality is be influenced by many factors and for more practical use are represented with classification systems like water quality classes (WQC), that will be focus our research using the fuzzy algorithm.

To analyse these type of ecological data, special analytic tools are developed that are much or less based on the classical statistical tools like variation, standard deviation to more complex statistical tools and computer algorithms. Research indicates that they widely used and have easy interpretation. Introduction of some basic algorithms and their implementation is presented in [2] and [3]. Development of more advance algorithms and methods for this application and use are introduced by [4, 5, 6, 7, 8]. In these papers, authors discuss the development of methods like CART, and its successors – bagging, boosted trees, random forest and other techniques in ecological studies. The newer methods and increasing research done by [9] and [10], have used these methods for eco-hydrological modelling and prediction in plant types of wet habitat.

Similar methods, especially the methods that are used for classification and regression, the well-known decision trees are applied for diatom classification of Lake Prespa. Lake Prespa models were obtained to detect the influence of the physic-chemical parameters on the diatom community [11]. Several models with different type of settings were applied on the algorithm to test the robustness and the classification accuracy used. Then these models were discussed by the biological expert that have collected the samples. After careful consideration of the models and the relevant biological knowledge at that time, some of the model obtained have confirmed the correct indication of the diatom property, and some of the produced knowledge about the diatoms was new and yet to be tested by the biological diatom experts [11] and [12]. Following these results, another method was used that increased the classification accuracy of the previously used method. This method included use of the algorithm for multi target modelling, thus predicting the influence of the several physic-chemical parameters on the diatoms. Nonetheless, some properties that were part of this algorithm like, not be robust on data change and not resisting on over-fitting that inherently were implied by the dataset and the nature of the algorithm. Some of these properties that were consider as disadvantage must be improved, and some of them can be overcome using the fuzzy theory. Previous research studies [13, 14] have proven that the fuzzy method used in this paper is more appropriate for this task. In this direction, the algorithm presented in this paper is improvement of this original algorithm for fuzzy modelling and can be used to extract the ecological information. The method in this paper is robust to over fitting, as it is shown in [13, 14]. The obtained results in a form of rules can be easily interpreted and compared with the known ecological diatom indicator references in the literature [15]. The last improvement of this algorithm against the other algorithm is that he has achieved higher classification accuracy. One of the reason is better, is that uses different fuzzy membership functions and similarity metrics to fuzzy the data. Several research papers have shown this [16]. In order to express the spatial space information embedded in the models from the measure data for each measuring station, we will present one model in GIS. The GIS models combine the classification property of the algorithm and the visualization power of the software. The GIS models graphically depict the low and high abundance values, which show the most optimal habitat of given diatom for all the physico-chemical measured parameters.

The rest of the paper is organized as follows: Section II provides the definitions for similarity and aggregations metrics. In Section III we briefly introduce the fuzzy membership function used. Section IV presents the diatoms abundance water quality datasets and the experimental setup, while the section V gives the experimental results and the interpretation for some of the model trees generated by the algorithm and the spatial fuzzy models. Finally, Section VI concludes the paper and research direction is outlined.

II. SIMILARITY AND FUZZY AGGREGATION METRICS

The classification diatom model is obtained by using similarity metric and fuzzy aggregation operator, which definition are presented in this section. The root mean square error (RMSE) of fuzzy sets A and B can be computed as:

$$Sim(A;B) = 1 - RMSE(A;B) = \sqrt{\frac{\sum_{i=1}^{n}\left(\mu_A(x_i) - \mu_B(x_i)\right)^2}{n}} \qquad (1)$$

,where x_i, i = 1, . . . , n, are the crisp discretized values, and $\mu_A(x_i)$ and $\mu_B(x_i)$ are the fuzzy membership values of x_i for *A* and *B*, that are two fuzzy sets defined on the universe of discourse *U*. The larger the value of *Sim* (*A*; *B*), the more similar *A* and *B* are. As $\mu_A(x_i)$, $\mu_B(x_i)$ [0, 1], $0 \le Sim\,(A;B) \le 1$ holds according to (1). This metric was proposed by [17].

The second step in the inducting the model is the operation between two fuzzy membership values. According fuzzy logic theory, the fuzzy aggregation are logic operators applied to fuzzy membership values or fuzzy sets. They have three sub-categories, namely t-norm, t-conorms, and averaging operators such as weighted averaging (WA) and ordered weighted averaging (OWA) [18]. In our experimental setup, we use the weighted operators, namely the weighted geometric operator (WG) and the ordered weighted geometric operator (OWG). The goal of the proposed fuzzy WG and OWG is to take into account the property of ratio measurements in the dataset [19]. The definition of the WG and the OWG operators as is follows:

A WG operator of dimension n is a mapping $R_n \rightarrow R$, that has an associated n-elements vector $w = (w_1, w_2, \ldots, w_n)^T$, $w_i \in [0, 1]$, $1 \le i \le n$, and that:

$$WG(a_1, a_2, \ldots, a_m) = \prod_{j=1}^{m} w_j a_j\,, \quad \sum_{i=1}^{m} w_i = 1 \qquad (2)$$

An OWG operator [17] of dimension n is a mapping $R_n \rightarrow R$, that has an associat-ed n-elements vector $w = (w_1, w_2, \ldots, w_n)^T$, $w_i \in [0, 1]$, $1 \le i \le n$, and

$$OWG^G(a_1, a_2, \ldots, a_m) = \prod_{i=1}^{m} c_i^{w_i} \quad \sum_{i=1}^{m} w_i = 1 \qquad (3)$$

where c_i $(a_1, a_2 \ldots, a_n)$ returns the i-th largest element of the collection $\{a_1, a_2 \ldots, a_n\}$.

The fundamental difference of OWG from WG aggregation operator is that the former does not have a particular weight - w_i associated for an element; rather a weight is associated with a particular ordered position of the element. Also important property of the OWG and WG operators compared with the OWA and the WA is the property of ration that we mention before mainly used in other research areas [20]. In order to compare the results from our previous research the geometric weight operator with simpler fuzzy membership functions we use compare the triangular, trapezoidal and Gaussian membership function with the sigmoidal membership function.

III. SIGMOID MEMBERSHIP FUNCTION

Using simpler membership functions, the process of generating fuzzy terms can be achieved using three evenly distributed MFs: trapezoidal, triangular and Gaussian. Because the diatom dataset have relationship with the physic-chemical parameters that is different from these distribution, another distribution is used to fuzzify the input of the algorithm. The definition of the evenly sigmoid MF, which in general is specified by three parameters, is represented with (4).

$$f(x;a;b) = \frac{1}{1 + e^{-a*(x-b)}} \qquad (4)$$

In (4) the parameter are constants, while the *b* parameter is located at the center of the curve. In this paper, we modify the (4), by taking the mean values of the given data range into account. In this way, each fuzzy term will reflect the very nature of the dataset and evenly distributed sigmoid MF in the entire range. And finally when all this changes are taken into account (5) mathematically represents the sigmoid MF as:

$$f(x;\mu;\sigma) = \frac{1}{1 + e^{-a*(x-\mu)}} \qquad (5)$$

In (5) the parameter *a* can get two values {1 and -1} and both of these values will be studied in this paper. Because of the smoothness and concise notation, the sigmoid MF, can be used for specifying fuzzy sets and ecological knowledge discovery and be used for wide range of different type of datasets.

IV. DATA DESCRIPTION AND EXPERIMENTAL SETUP

The datasets used in the experiments consist from 13 input parameters representing the TOP10 diatoms species (diatoms species that exist in Lake Prespa [21]) with their relative abundance per sample, plus the three WQCs for conductivity, pH and Saturated Oxygen. These measurements were made as a part of the TRABOREMA project [22]. The water quality classes are defined according the three physical-chemical parameters: Saturated Oxygen [23], Conductivity [24] and pH [23, 24] which are given in Table 1.

Among the input parameters, 10 are numerical parameters and the rest 3 are nominal with number of possible classes from 4 up to 6. We have made two variants of the method.

TABLE I. WQCs FOR THE PHYSICAL-CHEMICAL PARAMETERS

Physical-chemical parameters	Name of the WQC	Parameter range
Saturated Oxygen	*oligosaprobous*	SatO > 85
	β-mesosaprobous	70-85
	α-mesosaprobous	25-70
	α-meso / polysaprobous	10-25
pH	*acidobiontic*	pH < 5.5
	acidophilous	pH > 5.5
	circumneutral	pH > 6.5
	alkaliphilous	pH > 7.5
	alkalibiontic	pH > 8
	Indifferent	pH >9
Conductivity	*fresh*	Conductivity <20
	fresh brackish	Conductivity <90
	brackish fresh	90-180
	brackish	180-900

In order to produce models that compromise between complexity and performance we build two type of models. First type are the simple diatom model (SPT), which consist from 0 candidate trees and one low level tree with two different depths 5 – SPT5 and 10 - SPT10.

TABLE II. AVERAGE PREDICTION ACCURACY PER WQC (IN %)

Water Quality	Conductivity WQC – Average Prediction Accuracy (in)				
Type of experiment	*Triangular*	*Trapezoidal*	*Gaussian*	*Sigmoid (a= 1)*	*Sigmoid (a= -1)*
Train	76.15	76.15	76.15	**77.52**	75.23
Exp2	**73.70**	72.32	73.25	73.16	71.77
Classical Algo.	*C 4.5*	*kNN*	*Bagging C4.5*	*Boosted C4.5*	
xVal	65.60	66.51	63.30	63.76	
Water Quality	**pH WQC – Average Prediction Accuracy (in)**				
Type of experiment	*Triangular*	*Trapezoidal*	*Gaussian*	*Sigmoid (a= 1)*	*Sigmoid (a= -1)*
Train	61.47	62.84	**63.76**	61.93	61.01
Exp2	58.12	58.14	58.14	59.94	**60.39**
Classical Algo.	*C 4.5*	*kNN*	*Bagging C4.5*	*Boosted C4.5*	
xVal	54.73	47.26	53.23	56.22	
Water Quality	**Saturated Oxygen – Average WQC Prediction Accuracy (in)**				
Type of experiment	*Triangular*	*Trapezoidal*	*Gaussian*	*Sigmoid (a= 1)*	*Sigmoid (a= -1)*
Train	**64.68**	61.19	63.68	59.70	60.20
Exp2	58.50	55.50	59.00	55.50	56.00
Classical Algo.	*C 4.5*	*kNN*	*Bagging C4.5*	*Boosted C4.5*	
xVal	**62.13**	61.13	62.00	62.00	

And second, we induce models that consist from 2 candidate trees, 3 low level trees and depth equal to 5 – PT5 and 10 – PT10. For similarity definition we use RMSE similarity and the WG and OWG fuzzy aggregation operator. Later, comparison with other crisp classifiers is done with simple and general models with different depth (5 and 10).

The configuration of the experiments is set up as follows. 1) A simple fuzzification method based on three evenly distributed MFs including the modified sigmoid MF for each input variable is used to transform the crisp values into fuzzy values (Train); and 2) Standard 10-fold cross validation is used for testing the prediction performance accuracy of the built models (xVal). Table 2 shows results of the conducted experiments.

V. EXPERIMENTAL RESULTS

A. *Performance Analysis*

The sigmoid shaped MF outperforms 2 of the 3 diatoms WQCs. Both triangular and Gaussian function in combination of the two modified sigmoid MFs have obtained higher prediction accuracy. If we compare the fuzzy algorithm and compare with the classical classification algorithms, than we can see the only for the Saturated Oxygen WQC. Better classification accuracy by the fuzzy classification algorithm is achieved by sigmoid and triangular MF, for conductivity and pH respectively. It is important to note, that the classification accuracy for some of the models obtained by the fuzzy algorithm is some case is more than 5% accuracy. If we pay attention in the variants of the fuzzy algorithm, the simple PT variants have achieved better descriptive and predictive accuracy. Saturated Oxygen WQC gained low performance, because the shape of the MF is not suitable for this WQC. However, it remains in focus for our future research.

B. Interpretation of the diatom models

In the previous section, we have compared the performance of the algorithm, and base on the results in this section we present several fuzzy diatom models and their interpretation.

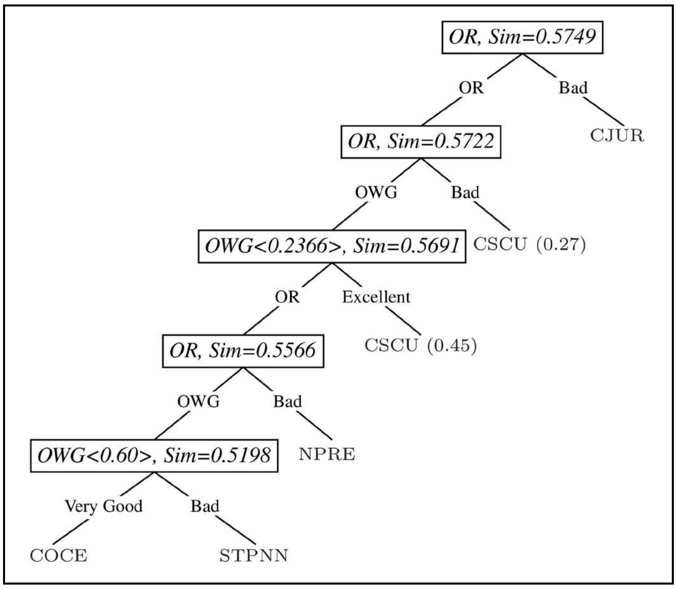


Fig. 1. Diatom model obtained for the *fresh* -Conductivity WQC

In order to achieve better interpretation and the compatibility with the Water Framework Directive, we restricted the number of fuzzy terms to 5. All the models that are presented use experimental setup 2. The diatom model shown in Fig. 1 can be converted into rule which is stated below.

Rules1: **IF** Conductivity WQ class is ***fresh*** **THEN** (*Cyclotella ocellata* (COCE) is **Very Good Indicator** OWG<0.6> *Staurosirella pinnata* (STPNN) is **Bad Indicator**) OR *Navicula prespanense* (NPRE) is **Bad Indicator** OWG <0.23> *Cavinula scutelloides* (CSCU) is **Excellent Indicator** OR CSCU is **Bad Indicator** OR CJUR is **Bad Indicator**. The model has highest similarity of 57.49%.

From Rule1 it can be easily seen that the two main diatoms CSCU and COCE, especially CSCU with higher abundant than the COCE diatom can be found in the water were Conductivity class is *fresh*. According the tree model, bad indicator or these diatoms cannot be found in *fresh* waters are STPNN, NPRE and CJUR diatoms. According the tree model, bad indicator or these diatoms cannot be found in *fresh* waters are STPNN, NPRE and CJUR diatoms.

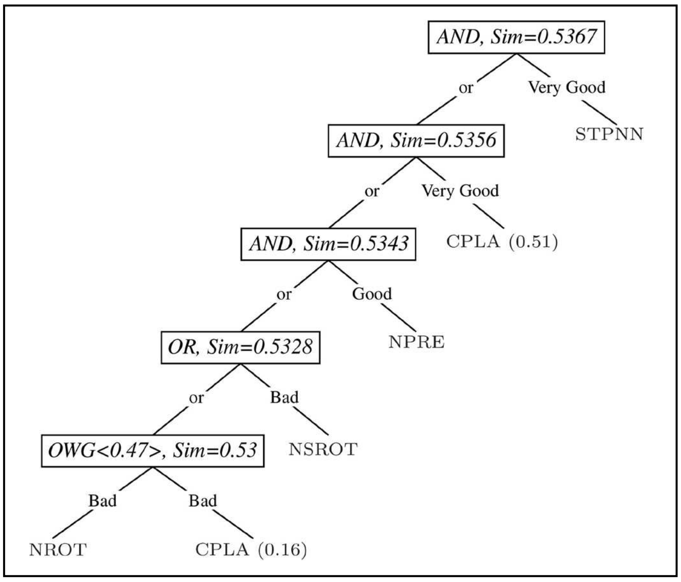


Fig. 2. Diatom model obtained for the *β-mesosaprobous* Saturated Oxygen

The rule induced from tree shown in Fig. 2 states:

Rule2: **IF** Saturated Oxygen WQ class is ***β-mesosaprobous*** **THEN** (*Navicula rotunda* (NROT) is **Bad Indicator** OWG<0.47> *Staurosirella pinnata* (CPLA) is **Bad Indicator**) OR *Navicula subrotundata* (NSROT) is **Bad Indicator** AND NPRE is **Good Indicator** AND CPLA is **Very Good Indicator** AND STPNN is **Very Good ad Indicator**. The model has highest similarity of 53.67%.

This model tree shows that three diatoms can be found to exist in ***β-mesosaprobous*** waters. According to the tree model, the CPLA and STPNN diatoms are more likely to be found with NPRE in these waters.

C. Spatial fuzzy diatom models

The diatom model given in Fig. 1, shows the leaf 4 as spatial fuzzy model (see Fig. 3) in order to investigate the spatial information. According to the model, the NPRE diatom can be relative good indicator for low level concentration of metals in locations L_4, L_6 and L_5 while other location high concentration of metals (Cu, Mg, Mn and Zn) especially Zn is suitable for this diatom. Eutrophication parameters (nitrogen and phosphorus) for almost all measuring station are low. According the model, not for all the stations can be seen a model that is because there is no data for them. Each leaf from the diatom model can be represented with GIS, in this way for all the diatoms can be found the indicating properties.

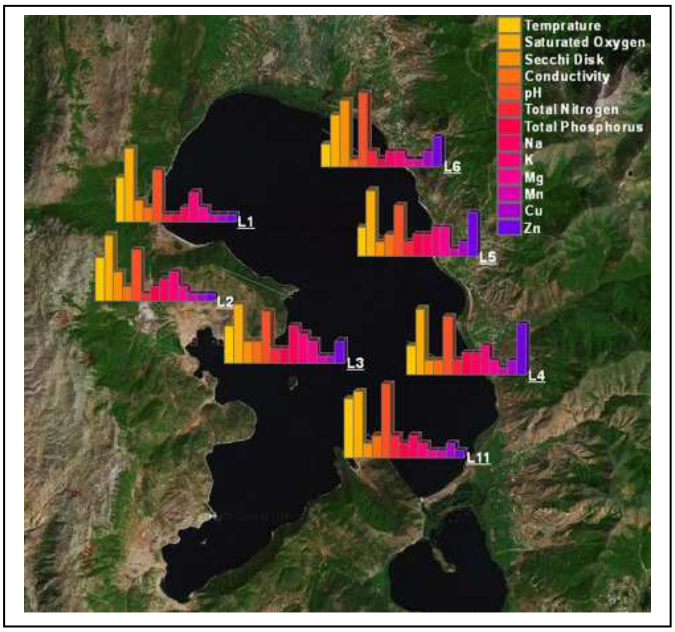


Fig. 3. Spatial fuzzy model for the NPRE diatom for *fresh* Conductivity WQ class.

VI. Conclusion

In this paper we have presented an algorithms, as we show in this research can be used for extraction of knowledge from measured dataset. The experiments on the WQC diatom dataset showed that provided sigmoid MF outperform models that use trapezoidal, triangular or Gaussian in terms of prediction accuracy for some WQC. Also, we compared the prediction accuracy between the proposed method and the ordinary crisp classification algorithms and showed improvement of the classification accuracy for some of the WQC datasets. The experimental evaluation of the experimental results show that the average prediction accuracy for the sigmoid MF is greater than the classical crisp classifiers. In terms of interpretability, the obtained models are easy interpretable and as the GIS models shown they can be used for spatial information representation.

Based on these results, we can conclude that the developing of more MFs is important to increase the classification accuracy of the models. Also, novel similarity metrics are also needed to increase the accuracy. This will lead to more spatial model that will express the important of the GIS modelling.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

References

[1] P. Legendre, H.J.B. Birks, “Clustering and partitioning:, *In: Tracking Environmental Change Using Lake Sediments" vol. 5, Data Handling and Numerical Techniques*, ed. H. J. B. Birks, A. F. Lotter, S. Juggins, & J. P. Smol, Dordrecht: Springer, 2011.

[2] G. De’ath, K.E. Fabricus, “Classification and regression trees: a powerful yet simple technique for ecological data analysis,” *Ecology,* vol. 81, pp. 3178-3192, 2000.

[3] A.H. Fielding, “Cluster and Classification Techniques for the Biosciences,” *Cambridge: Cambridge University Press*, 2007.

[4] G. De’ath, “Multivariate regression trees: a new technique for modelling species–environment relationships,” *Ecology*, vol. 83, pp. 1105-1117, 2002.

[5] G. De’ath, “Boosted trees for ecological modelling and prediction,” *Ecology,* vol. 88, pp. 243-251, 2007.

[6] A.M, Prasad, L.R. Iverson, A. Liaw, “Newer classification and regression tree techniques: bagging and random forests for ecological predictions,” *Ecosystems,* vol. 9, pp. 181-199, 2006.

[7] D.R. Cutler, T.C. Edwards, K.H. Beard, “Random forests for classification in ecology,” *Ecology,* vol. 88, pp. 2783-2792, 2007.

[8] J. Elith, J.R. Leathwick, T. Hastie, “A working guide to boosted regression trees,” *Journal of Animal Ecology*, vol. 77, pp. 802-881, 2008.

[9] J. Peters, B. De Baets, N.E.C. Verhoest, “Random forests as a tool for ecohydrological distribution modelling,” *Ecological Modelling,* vol. 207, pp. 304-318, 2007.

[10] E. F. Stroemer, and J. P. Smol, *The diatoms: Applications for the Environmental and Earth Sciences* 2nd Edition. Cambridge University Press, Cambridge, pp.23-47 (2010).

[11] A. Naumoski, and K. Mitreski, "Diatoms classification with weighted averaging fuzzy operators for eutrophication prevention," In P. Golinska et al. (Eds.), *Information Technologies in Environmental Engineering – Environemtal Science and Engineering* 3, DOI: 10.1007/978-3-642-19536-5_4, Springer-Verlag Berlin Heidelberg, pp. 49-60, 2011.

[12] A. Naumoski, and K. Mitreski, "Classifying diatoms into trophic state index classes with novel classification algorithm, "Procedia Environmental Sciences 2, pp. 1124-1138, 2010.

[13] A. Naumoski, G. Mirceva, and K. Mitreski, "A novel fuzzy based approach for inducing diatom habitat models and discovering diatom indicating properties," *Ecological Informatics* 7(1), pp. 62-70, 2012.

[14] A. Naumoski, and K. Mitreski, "Novel algorithm for diatom classification in Lake Prespa using log-normal distribution," *International Journal of Ecohydrology & Hydrobiology*, 11(1-2), pp. 23-34, 2011.

[15] M.H. Van Dam, “A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands,” *Netherlands Journal of Aquatic Ecology* vol. 28(1), pp. 117-133, 1994.

[16] A. Naumoski, K Mitreski, "Novel fuzzy operator for classification of diatoms in ecological water quality classes," *In ICT Innovations 2013*, Vladimir Trajkovic and Anastas Mishev (Eds.), ISSN 1857-7288, pp. 72-81, 2013.

[17] Z.H. Huang, T.D. Gedeon, and M, Nikravesh, M, “Pattern Trees Induction: A New Machine Learning Method.” *IEEE Transaction on Fuzzy Systems,* vol. 16, no. 3, pp. 958-970, 2008.

[18] R.R. Yager, “On ordered weighted averaging aggregation operators in multicriteria decision making,” *IEEE Transactions on Systems, Man, and Cybernetics,* vol, 18, pp. 183-190, 1988.

[19] J. Azcel, C. Alsina.,”Synthesizing judgements: functional equations approach,” *Mathematical Modelling,* vol. 9, pp. 311-320, 1987.

[20] J. Azcel, C. Alsina, “Procedures for synthesizing ratio judgements,” *Journal of Mathematical Psychology*, vol. 27, pp. 93-102, 1983.

[21] Z. Levkov, S. Krstič, D. Metzeltin, and T. Nakov, “Diatoms of Lakes Prespa and Ohrid (Macedonia). *Iconographia Diatomologica*. vol. 16, pp. 603, 2006

[22] TRABOREMA Project - WP3: EC FP6-INCO project no. INCO-CT-2004-509177, 2005-2007.

[23] K. Krammer, and H. Lange-Bertalot, “Die Ssswasserflora von Mitteleuropa 2: Bacillariophyceae. 1 Teil,”*Stuttgart: Gustav Fischer-Verlag*, pp. 876, 1986.

[24] A. Van Der Werff, and H. Huls, *Diatomeanflora van Nederland.* Abcoude - De Hoef, 1957, 1974.

Collaborative Personalized Healthcare Algorithm: Development and Evaluation

Igor Kulev*, Elena Vlahu-Georgievska†, Saso Koceski* and Vladimir Trajkovik‡

*Faculty of Computer Science and Engineering
University "Ss Cyril and Methodious", Skopje, Macedonia
{igor.kulev, trvlado}@finki.ukim.mk
†Faculty of Administration and Information Systems Management
University "St.Kliment Ohridski", Bitola, Macedonia
elena.vlahu@uklo.edu.mk
‡Faculty of Computer Science
University "Goce Delcev", Stip, Macedonia
koceski@yahoo.com

Abstract—**Providing patients with convenient health facilities at a low cost has always been a great challenge for health service providers. Moreover, the fast changing life style of the modern world and the problem of aging society pose an urgent need to modernize such facilities. The emphasis has to be paid on providing health monitoring in out-of-hospital conditions for elderly people and patients who require regular supervision, particularly in remote areas. One of the health parameters that need to be controlled is blood sugar. Out of control blood sugar levels can lead to serious short term problems such as hypoglycemia, hyperglycemia, or diabetic ketoacidosis. The recommendation algorithm evaluated in this paper incorporates collaboration and classification techniques in order to generate recommendations and suggestions for the physical activities that the users should carry out in order to improve their health. In this paper we have shown how the proposed algorithm can be implemented in real-world situations and we have successfully evaluated it using generic data based on realistic modeling of food intake influence over the blood glucose level.**

Index Terms—**health care, recommendation algorithm, evaluation, modeling, parameter regulation.**

I. Introduction

Terms "telemedicine" or "telehealth" encompass a wide range of services that use information and communication technologies to either provide care, or support care provided electronically over a geographic distance. In that way, telehealth can save lives, reduce costs and improve patient access to care.

Information and communication technologies has become indispensable to health workers, as the volume and complexity of knowledge and information have outstripped the ability of health professionals to function optimally without the support of information management tools [1]. So, there is an urgent need for information and communication tools that can gather information from multiple sources and provide a new point of view of human health. Information and communication technologies make it possible to bridge the gap and time barriers in the flow of health information and knowledge, allowing every involved part in the health process to have access to the information. This approach provides the knowledge of the individual to contribute effectively to the improvement in human health. But also, helps the collective knowledge effectively to solve health problems on individual level and globally.

A huge implication for health has the vision of the ubiquitous computing. These technologies support systems that provide services for continuous health monitoring of the patients and communication with health centers [2]. Developing an information system that offers continuous monitoring of health data, food intake and patients' environment is very important for both, healthcare providers and patients. Such systems could also support the process of decision making by searching through large amounts of health data and facts, classifying them and identifying issues that are directly relate to a given medical condition. So, they could offer citizens to be directly involved in their health care, providing information that will assist in making decisions about their own health [3]. Moreover, patients will have a greater role in the decision making processes related to their health as they could be empowered with the ability to gain access and manage information that fits with their personalized needs, and ultimately, to shape their health as a reflection of the whole community.

COHESY [4] uses bionetwork, mobile, web and broadband technologies. Bionetwork is implemented by various body sensors that measure the value of the patients' health parameters. Broadband mobile technology provides movements of electronic care environment easily between locations and internet-based storage of data allows moving location of support. Different sensors could be connected to a mobile device (for example smartphone) and an application installed on the device could read the sensor data. Sensors are considered as relevant source of data. They are also used to confirm users' feedback (e.g. description of activity). The use of a social network allows communication between users with same or similar condition and exchange of their experiences. This system model has simple graphical interfaces that provide easy use and access not only for the young, but also for elderly users. It has more purpose and includes use by multiple categories of users (patients with different diagnoses). Some of its advantages are scalability and ability of data information storing when communication link fail. COHESY is interoperable system that allow data share between different systems and databases.

The recommendation algorithm [5] is part of the social network in COHESY. This algorithm uses the data read by the bionetwork, the data about the user's physical activities (gathered by the mobile application), user's medical record (obtained from clinical centre) and the data contained in the user profile on the social network (so far based on the knowledge of the social network). The main purpose of this algorithm is to find the dependency of the users' health condition and physical activities they perform. So, the algorithm generates recommendations for the physical activities that the users should carry out in order to improve their health. In order to generate recommendations the algorithm incorporates collaboration and classification techniques. The classification algorithms are used on datasets from the health history of users for grouping the users based on their similarity. Use of classified data when generating the recommendation provides more relevant recommendations because they are enacted on knowledge for users with similar medical conditions and reference parameters.

In this paper we are using simulations on generated data to see how different types of activities are affecting the accuracy of the algorithm. The generated data consists of information about activities (most important are date and type of activity) and information about the measurements (date of measurements and parameter value). On the basis of the activities and measurements, our recommendation algorithm should determine which type of activity has bigger influence on the change of the blood glucose levels. This experimental result can lead us to explore the domain where the algorithm would have the best utilization.

In the next section experimental methodology and result will be discussed. The third section is the conclusion of the paper.

II. Evaluation of the algorithm

In this evaluation we want to determine which type of activity has bigger influence on the change of the blood glucose levels. We want to evaluate our algorithm and to calculate its accuracy in different phases of the simulation. We would also like to explore how the accuracy changes when we change some of the parameters in our simulator.

A. Methodology

The data used in our experiments is provided by a simulator. We did a simulation where we simulate the influence of the food intake and the activities on the change of the blood glucose level. We assume that after food intake, first there is big increase and after that there is small decrease of the blood glucose level. It is important that in absence of activities, food intake should increase the value of the blood glucose after sufficiently long time. We simulate two food intakes in each day: breakfast and lunch. Lunch has bigger influence on the blood glucose level than breakfast. This can be seen on Fig. 1.

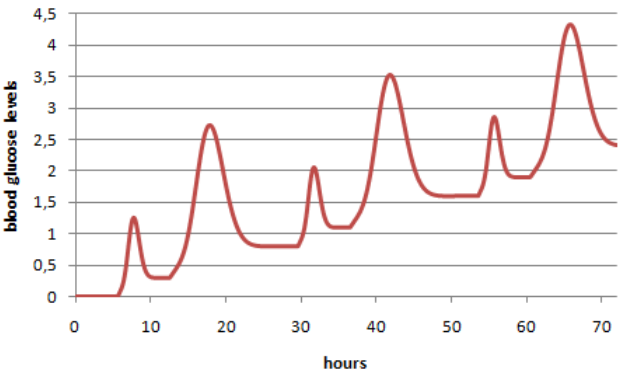


Figure 1. Blood glucose levels in three days if we assume that there are two food intakes each day (breakfast and lunch) and there are no activities.

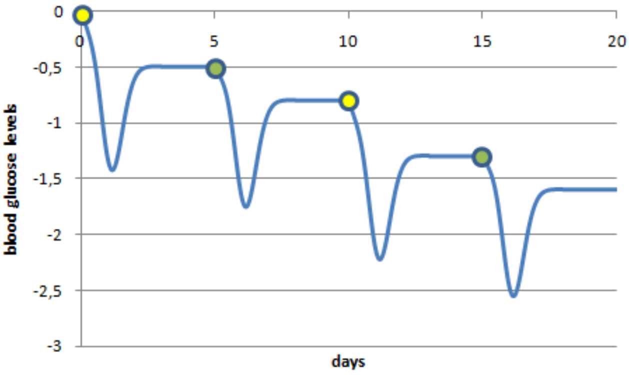


Figure 2. Fig. 2. Blood glucose levels if we assume that there are 4 activities of two types. The first type (yellow) has larger influence on the parameter value and the second type (green) has smaller influence on the parameter value, although both types of activities have the same peak value.

Activities have opposite effect on the change of the parameter value. In our simulation we use two types of activities. Both types of activities decrease the blood glucose levels. On Fig. 2 we can see the impact of the activities on the blood glucose levels when there is no food intake.

Our simulator has 11 parameters. The initial parameter values are chosen by observation and they are used as reference parameters. Changing the default parameter values and repeating the experiments which are described later should not change the conclusions that will be made.

- p_b - peak of the breakfast influence (default: 1)
- c_b - change of the blood glucose levels after sufficiently long time caused by breakfast (default: 0.3)
- p_l - peak of the lunch influence (default: 2)
- c_l - change of the blood glucose levels after sufficiently long time caused by lunch (default: 0.5)
- p_f - peak of the first type of activity (default: -1)

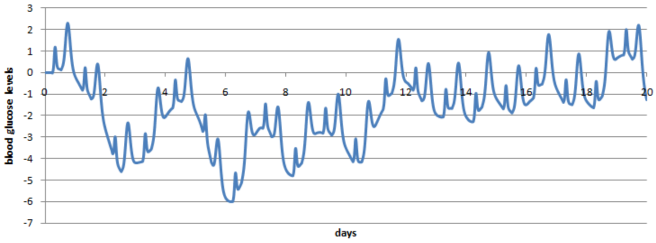


Figure 3. Change of the blood glucose levels during one simulation.

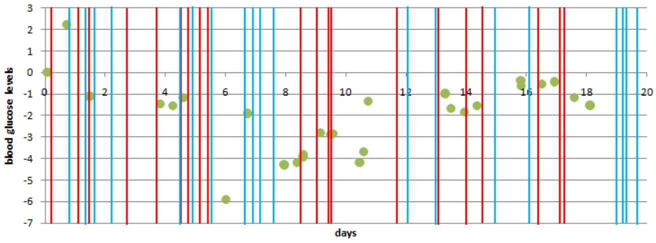


Figure 4. On the graph we can see the data that is provided to the recommendation algorithm. The moments when the first type of activity has occurred are denoted with blue vertical lines. The moments when the second type of activity has occurred are denoted with red vertical lines. The measurements are denoted with green circles.

- c_f - change of the blood glucose levels after sufficiently long time caused by the first type of activity (default: -0.5)
- p_s - peak of the second type of activity (default: -1)
- c_s - change of the blood glucose levels after sufficiently long time caused by the second type of activity (default: -0.3)
- L - Length of the simulation in days (default: 20 days)
- M - Number of measurements (default: 30)
- N - Number of activities of each type

We have decided that we will have the same number of activities of each type and that the expected change of the blood glucose levels after sufficiently long time is zero when there we include the food intakes and the activities in our simulation. We can calculate N in this way:

$$N = \frac{-L \cdot (c_b + c_l)}{c_f + c_s} \tag{1}$$

For the default values of the parameters $N = 20$. We generated the activities and measurements at random moments. The change of the blood glucose levels during one simulation is shown on Fig. 3. The data that is provided to the recommendation algorithm (activities and measurements) is shown on Fig. 4.

B. Results and analysis

In different scenarios we have changed the values of the parameters and we have observed how the accuracy of our recommendation algorithm changes as the time of the simulation increases. For each scenario we have performed 10,000 simulations. In every simulation we have used the recommendation algorithm at different phases to find the type of activity that causes bigger decrease of the blood glucose levels.

In the simulations with the default parameter values (blue line on Fig. 5), we can observe high accuracy at the start of the simulations (80%). This is expected because at the start of the simulations, there are fewer activities that have influence on the blood glucose levels and this is why it is easier for the recommendation to recognize the type of activity which causes bigger decrease of the parameter value. As the time goes by, the accuracy decreases. This is because the parameter value is influenced by more activities as the time goes by, so it becomes more difficult for the recommendation algorithm to determine which activity had the biggest effect on the parameter value. After five days the function reaches its minimum. In the following days the accuracy of the algorithm increases because there is bigger amount of data available. In the end of the simulations, the accuracy reaches 58%. We compare the results of this experiment with the results from 7 types of simulations with changed parameter values:

- $c_f = -0.9, c_s = -0.7$ - we increase the change of the blood glucose levels after sufficiently long time caused by both types of activity by a constant (Fig. 5a)
- $c_b = 0.7, c_l = -0.9$ - we increase the change of the blood glucose levels after sufficiently long time caused by breakfast and lunch by a constant (Fig. 5b)
- $c_f = -0.9, c_s = -0.3$ - we increase the change of the blood glucose levels after sufficiently long time caused by both types of activity and the increase for the first type of activity is bigger (Fig. 5c)
- $t_i = 6$ - we increase the length of the period in which the stable influence of the activities is achieved (Fig. 5d)
- $p_b = 2, p_l = 3$ - we increase the magnitude of the peaks of the breakfast influence and the lunch influence by a constant (Fig. 5e)
- $p_f = -2, p_s = -2$ - we increase the magnitude of the peaks of both types of activities by a constant (Fig. 5f)
- $c_f = -2, c_s = -1.5$ - we increase the magnitude of the peaks of both types of activities and the increase of the magnitude of the peak of the first type of activity is bigger (Fig. 5g)

From Fig. 5 we can see that the only case when there is significant difference between the reference results and the results obtained from the simulations with changed parameter values is when there is bigger difference between the changes of the blood glucose levels after sufficiently long time caused by the first and the second type of activity (Fig. 5c). In this case, the algorithm reaches 80% accuracy in the end of the simulations. These results were as expected. However, it is surprising that the results from the other six experiments were very similar to the results from the reference experiment. From Fig. 5a and Fig. 5b we can conclude that if we increase the values of c_f, c_s, c_b and c_l by a constant, then the accuracy of the algorithm does not change. From Fig. 5d we can conclude

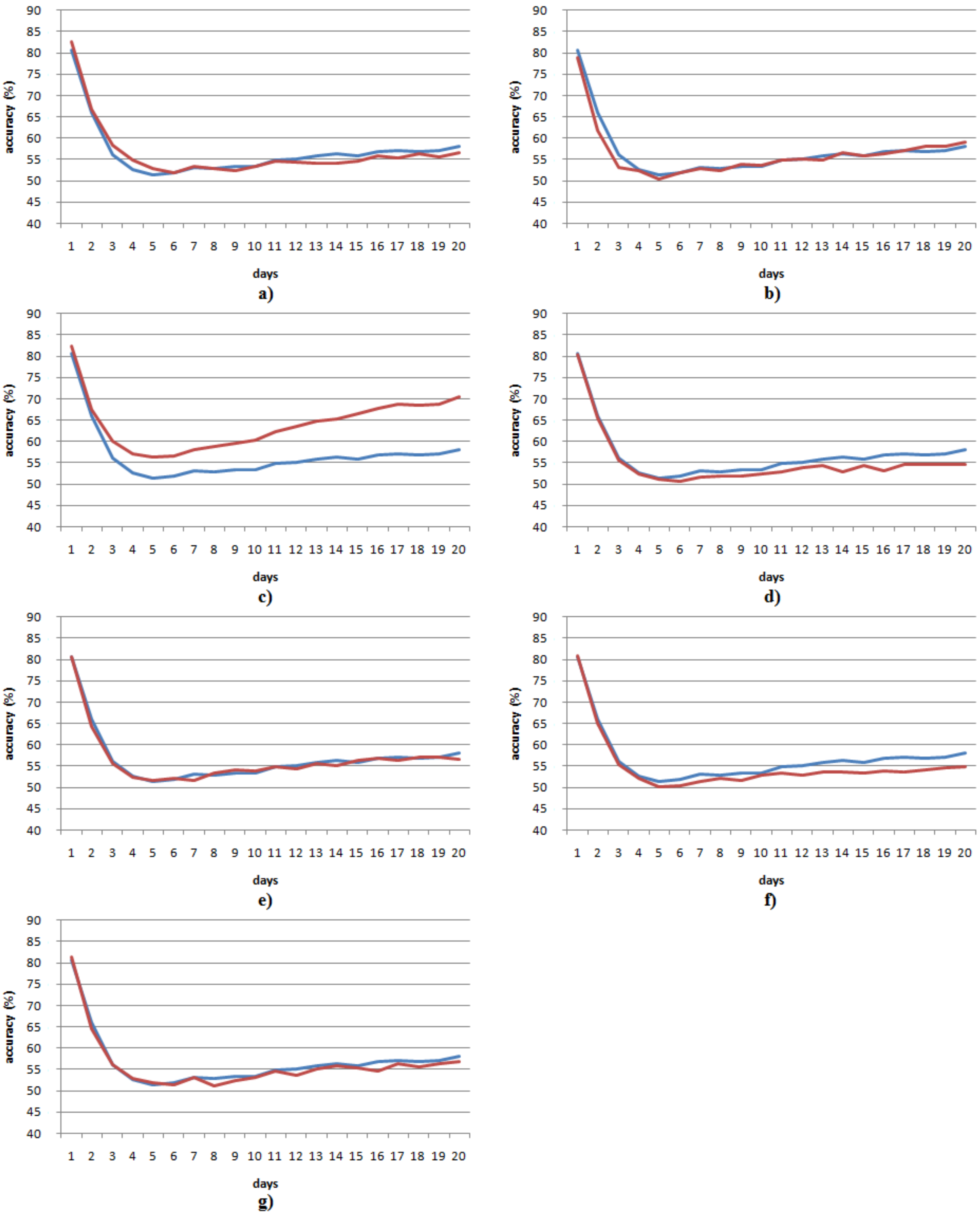


Figure 5. Accuracy as a function of the time of simulation. Blue line represents the results obtained from the simulations with default parameter values. Red line represents the results obtained from simulations with changed parameter values: a) $c_f = -0.9$, $c_s = -0.7$ b) $c_b = 0.7$, $c_l = 0.9$ c) $c_f = -0.9$, $c_s = -0.3$ d) $t_i = 6$ e) $p_b = 2$, $p_l = 3$ f) $p_f = -2$, $p_s = -2$ g) $p_f = -2$, $p_s = -1.5$.

that if we increase t_i we will obtain lower accuracy. From Fig. 5e we can conclude that if we increase the values of p_b and p_l by a constant the accuracy is not changed. However, if we decrease the values of p_f and p_s by a constant, we obtain lower accuracy. On the other side, we decreased the values of p_f and p_s by different amounts and we observed that in this case the accuracy is not changed (Fig. 5g). This means that the magnitude of the peak has considerable effect on the accuracy of the recommendation algorithm. Better results would be achieved if $p_f - p_s > T$ and $c_f - c_s > T$ or $p_f - p_s < T$ and $c_f - c_s < T$ where T is some threshold.

III. Conclusion

In this paper we have presented a method that could be used to regulate blood glucose levels by recommending relevant activities. We have introduced a simulation model in which food intake and two types of activities have influence on the blood glucose levels in a specific way. We have shown that when there are few activities, the recommendation system could easily determine the needed activity. Also, bigger accuracy is achieved when there is a lot of data. When the amount of data is not sufficiently small, the algorithm shows its worst performance. We have tested how the accuracy changes when we change the values of the simulation parameters. The results show that the accuracy is larger if one of the activities has bigger peak and causes more significant change of the blood glucose levels after sufficiently long time than the other one. In our future work, we could make simulations with different mathematical models for the activity influence.

Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University.

References

[1] S. Y. Kwankam, “What e-health can offer,” *Bulletin of the World Health Organization*, vol. 82, no. 10, pp. 800–802, 2004.

[2] E. Sillence, L. Little, and P. Briggs, “E-health,” in *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 2*. British Computer Society, 2008, pp. 179–180.

[3] A. Cawsey, F. Grasso, and C. Paris, “Adaptive information for consumers of healthcare,” in *The adaptive web*. Springer, 2007, pp. 465–484.

[4] I. Kulev, E. Vlahu-Gjorgievska, S. Koceski, and V. Trajkovik, “Collaborative system for prevention of increased blood sugar level,” in *Proceedings of the 9th International Conference for Informatics and Information Technology*, 2012, pp. 115–119.

[5] I. Kulev, E. Vlahu-Gjorgievska, V. Trajkovik, and S. Koceski, “Development of a novel recommendation algorithm for collaborative health-care system model.” *Computer Science & Information Systems*, vol. 10, no. 3, 2013.

Protein function prediction using semantic hierarchical clustering algorithm

Ilinka Ivanoska, Kire Trivodaliev, Slobodan Kalajdziski
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Department of Intelligent Systems
Skopje, Macedonia
{ilinka.ivanoska, kire.trivodaliev, slobodan.kalajdziski}@finki.ukim.mk

Abstract—**The proposed protein function prediction methods are mostly based on sequence or structure protein similarity and do not take into account the semantic similarity extracted from protein knowledge databases such as Gene Ontology. Many studies have shown that identification of protein complexes or functional modules can be effectively done by clustering protein interaction network (PIN). A significant number of proteins in such PIN remain uncharacterized and predicting their function remains a major challenge in system biology. In this paper we present a "semantic driven" clustering approach for protein function prediction by using both semantic similarity metrics and the whole network topology of a PIN. We apply hierarchical clustering combined with several semantic similarity metrics as a weight factor in the distance-clustering matrix. Protein functions are assigned based on cluster information. Results reveal improvement over standard non-semantic similarity metric.**

Keywords— ***semantic similarity, protein function prediction, Gene Ontology, hierarchical clustering***

I. Introduction

Nowadays, most of the similarity-based methods for determining protein function rely on protein's sequence or structure. Unfortunately, the big drawback of these methods is that structure/sequence similarity is not directly related to the protein function, since proteins with significant structure/sequence similarity can have different functions. Furthermore, proteins with different ancestors and no significant sequence similarity can have the same function, due to evolution.

One of the most important challenges of molecular biology is finding a method for extracting protein function and protein similarity knowledge, consisted in the great amount of protein and genome data in well-known protein databases. An important breakthrough in protein annotation is the creation of the Gene Ontology (GO)[1], the most famous bio-ontology; structured and controlled vocabulary for describing gene and protein products. The GO, structured as a directed acyclic graph (DAG), defines a set of terms used for protein annotation. The GO-annotated interacting proteins can be used as a fertile basis for performing semantic driven protein comparison. This type of comparison is called semantic similarity, and is based on the structure of the GO and the relations between its terms, focusing on the semantic similarity between the terms themselves. It is still not clear which is the best way to calculate semantic similarity considering the current bio-ontologies, but several metrics have been proposed to calculate protein semantic similarity in the context of the GO[1],[2].

A protein interaction network (PIN) consists of nodes representing proteins, and edges representing interactions between proteins. Such networks are stochastic as edges are weighted with the probability of interaction. There is more information in a PIN compared to sequence or structure alone. A network provides a global view of the context of each gene/protein. Hence, our computational function prediction is characterized by the use of a protein's interaction context within the network to predict its functions.

It has been shown that clustering PINs is an effective approach to understand the relationship between the organization of a network and its function [3]. Clustering in PIN is to group the proteins into sets (clusters) that demonstrate greater similarity among proteins in the same cluster than in different clusters. Since biological functions can be carried out by particular groups of genes and proteins, dividing networks into naturally grouped parts (clusters or communities) is an essential way to investigate some relationships between the function and topology of networks or to reveal hidden knowledge behind them.

There are many clustering techniques proposed, but usually standard clustering algorithms are the most effective. In this paper we use hierarchical [4] clustering algorithm for standard protein clustering without the use of PIN, and for graph clustering with the use of PIN. Semantic similarity is added to these clustering techniques as a distance metric in the distance matrix. The aim is to present our work for evaluating the semantic similarity metrics and presenting a new system for protein function prediction by the use of this clustering algorithm based on semantic similarity. For a given protein the system can determine similar proteins based on functional (semantic) similarity, and our goal is to see the impact of semantic similarity metrics on determining protein function.

In section 2 we present a related work of the existing semantic similarity metrics that will be used, while section 3 will give the proposed system architecture for protein function prediction based on the semantic similarity metrics and the whole network topology using the semantic driven hierarchical clustering algorithm. Section 4 presents experimental results and a discussion of the way that semantic similarity metrics influence the prediction process. Finally, section 5 concludes the paper.

II. Overview of semantic similarity metrics

Several approaches are available to quantify semantic similarity between terms or annotated entities in an ontology represented as a DAG such as GO. There are essentially two types of methods for comparing terms in a graph-structured ontology: edge-based, that use the edges and their types as data source; and node-based, in which the main data sources are the nodes and their properties.

Edge-based approaches are based mainly on counting the number of edges in the graph path between two terms. The most common technique is the distance, that selects either the shortest path or the average of all paths, when more than one path exists. This technique gives a metric of the distance between two terms, which can be easily converted into a similarity metric. While this approach is intuitive, terms at the same depth do not necessarily have the same specificity, and edges at the same level do not necessarily represent the same semantic distance, therefore in this paper we do not take these metrics into account.

A. *Node-based metrics*

Node-based approaches are mainly based on comparing the properties of the terms involved, which can be related to the terms themselves, their ancestors, or their descendants. The most commonly used concept in these approaches is information content (IC), which gives a measure how specific and informative a term is. The IC of a term *c* is defined as the negative log likelihood *–log p*(*c*), where *p*(*c*) is the probability of occurrence of *c* in a specific knowledgebase, being normally estimated by its frequency of annotation. The use of IC is important because it is more probable (and less meaningful) that two gene products share a commonly used term than an uncommonly used term.

Four most common node-based semantic similarity metrics include Resnik's, Lin's, Jiang and Conrath's and Relevance metric. They were originally developed for the WordNet, and then applied to GO. The definition of Resnik metric [6] is the similarity between two terms as the IC of their most informative common ancestor (MICA) as the following:

$$sim_{\mathrm{Res}}(c_1, c_2) = IC(c_{MICA}) \tag{1}$$

Resnik's metric does not take into account how distant are the terms from their common ancestor. To consider that distance, Lin's metric [6] is defined as:

$$sim_{Lin}(c_1, c_2) = \frac{2 * IC(c_{MICA})}{IC(c_1) + IC(c_2)} \tag{2}$$

Jiang and Conrath's metric [7] is based on Resnik's metric considering the IC of the two terms compared as in Lin's metric. It is defined as follows:

$$sim_{JC}(c_1, c_2) = 1 - IC(c_1) + IC(c_2) - 2 * IC(c_{MICA}) \tag{3}$$

Another metric used is the relevance similarity metric [8], which is based on Lin's metric, but uses the probability of annotation of the MICA as a weighting factor to provide graph placement:

$$sim_{\mathrm{Rel}}(c_1, c_2) = sim_{Lin}(c_1, c_2) * (1 - p(c_{MICA})) \tag{4}$$

B. *Hybrid metrics*

Furthermore, there are hybrid methods that combine the two types of methods for semantic similarity, and they give weights to the GO nodes or edges according to their type. In this paper, for the purpose of protein function prediction, we use two semantic similarity hybrid metrics: Wang's [9] and Shortest path [10] semantic similarity metric.

In Wang's metric [9] each edge is given a weight according to the type of developed relationship. For a given term and its ancestor, a semantic contribution of the ancestor to the term is defined, as the product of all edge weights in the "best" path from the ancestor to the term, where the "best" path is the one that maximizes the product. Semantic similarity between two terms is then calculated by summing the semantic contributions of all common ancestors to each of the terms and dividing by the total semantic contribution of each term's ancestors to that term.

Shortest Path semantic similarity metric [10] uses a shortest path algorithm between terms in GO and gives weights to the terms as a reciprocal value of their IC. The distance between two terms is given by:

$$dist_{SP}(c_1, c_2) = \frac{\arctan\left(\sum_{t_1 \in path_1} \frac{1}{IC(t_1)} + \sum_{t_2 \in path_2} \frac{1}{IC(t_2)} \right)}{\pi / 2}$$

(5)

where *path₁(path₂)* is the shortest path between term $c_1(c_2)$ with its MICA, and $t_1(t_2)$ are the terms on $path_1(path_2)$. Semantic similarity is defined as

$$sim_{SP}(c_1, c_2) = 1 - dist_{SP}(c_1, c_2) \tag{6}$$

III. Protein function prediction system architecture

In our developed approach function prediction process is consisted of few steps: preprocessing; semantic matrix formulation; protein clustering and function prediction; and results evaluation. Fig. 1 shows the developed protein function prediction system architecture using semantic clustering algorithm. The preprocessing step [11] is made to get a highly reliably dataset. Following, semantic similarity between each protein pair is computed to formulate the dataset semantic similarity matrix. This means that we use the semantic similarity as a weighting factor while computing protein distance.

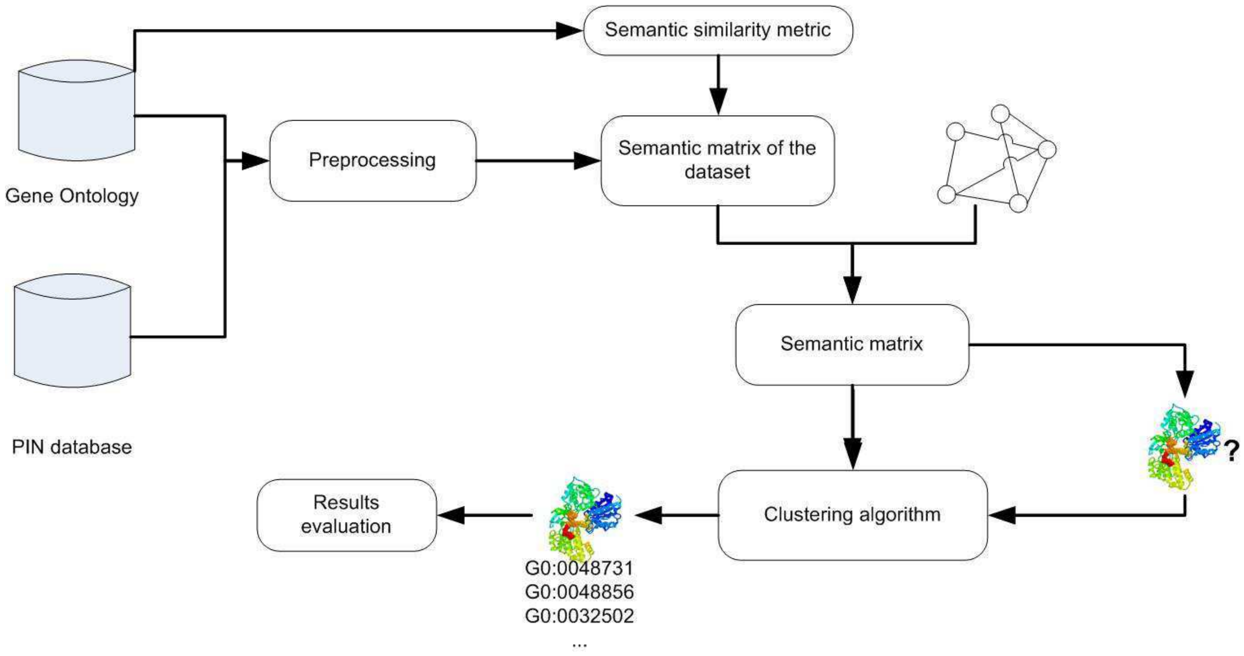


Fig. 1. Protein function prediction system architecture using semantic clustering algorithm.

We consider two scenarios: protein function prediction using standard hierarchical clustering without the use of PIN, and protein function prediction using PIN graph hierarchical clustering. In the former scenario, each protein pair semantic similarity matrix is an input to the hierarchical clustering algorithm as a distance matrix.

In the latter, the graph representing the PIN is weighted with the semantic similarity matrix where the weight shows the semantic protein distance (or the probability of interaction between protein pairs). The resulting semantic similarity matrix is an input to the graph hierarchical clustering algorithm.

After the clustering, we set up a strategy for annotating a query protein with the adequate functions according to the functions of the proteins in its cluster. Each function is ranked by its frequency of appearance as an annotation for the proteins in the cluster. This rank is calculated by (7) and it is then normalized in the range from 0 to 1.

$$f(j)_{j\in F} = \sum\nolimits_{i\in K} z_{ij} \tag{7}$$

where F is the set of functions present in the cluster K, and

$$z_{ij} = \begin{cases} 1, if \quad protein \quad i \quad from \quad K \quad has \quad function \quad j \quad from \quad F \\ 0, otherwise \end{cases} \tag{8}$$

In our proposed approach the previously explained Resnik's, Lin's, Relevance, JC's, Wang's and Shortest Path metrics are used in the SS matrix, and therefore, evaluated with the semantic driven protein function prediction algorithm.

IV. Results and Discussion

For the needs of this paper the dataset and the PIN data are compiled, pre-processed and purified from a number of established datasets, like: DIP, MIPS, MINT, BIND and BioGRID. The used dataset is believed to be highly reliable and consists of 2502 proteins from the interactome of the baker's yeast has 12708 interactions between them and are annotated with a total of 888 functional labels [11]. For the purposes of evaluating the proposed methods, the largest connected component of this dataset is used, which consists of 2146 proteins.

Each protein in the dataset is streamed through the prediction process one at a time as a query protein. The query protein is considered un-annotated, that is we employ the leave-one out method. Each of the algorithms works in a fashion that ranks the "proximity" of the possible functions to the query protein. The ranks are scaled between 0 and 1 as explained in 3. The query protein is annotated with all functions that have rank above a previously determined threshold ω. For example, for ω=0, the query protein is assigned with all the function present in its cluster. We change the threshold with step 0.1 and compute numbers of true-positives (TP), true-negatives (TN), false-positives (FP) and false-negatives (FN). For a single query protein we consider the TP to be the number of correctly predicted functions, and for the whole PIN and a given value of ω, the TP number would be the total sum of all single protein TPs. To compare performance between different algorithms we use standard measures as sensitivity (9) and false positive rate (10). Experiments were performed with different number of clusters and cutting; however, we conclude that the cutting in hierarchical clustering does not affect the results with different semantic similarity metrics.

$$sensitivity = \frac{TP}{TP+FN} \tag{9}$$

$$fpr = \frac{FP}{FP+TN} \tag{10}$$

For the purpose of comparing semantic similarity metrics with other non-similarity metrics we apply cosine similarity between proteins for the experiments that don’t involve PIN and minimal hop distance between proteins for those that involve the PIN.

The experiments were made for protein function prediciton using hierarchical clustering of the dataset with and without PIN. It is evident from table 1, that there is nonhomogeneity in the results with different semantic similarity metrics. The sensitivity for ω =0 varies from 65% for Resnic'c metric to 99% for Lin's and JC's metrics. In addition, the false positive rate varies from metric to metric. Table 1 shows that in the case of protein function prediction based on hierarchical clustering without PIN, the shortest path semantic similarity metric gives a decrease in the false positive rate of 55% compared to the standard non-semantic cosine metric, and a highest AUC value. Therefore, here we conclude that using a hybrid metric without PIN, is the best option.

TABLE I: PROTEIN FUNCTION PREDICTION RESULTS BASED ON SEMANTIC HIERARCHICAL CLUSTERING WITHOUT PIN

M	ω =	0	0.1	0.2	0.3	0.5	0.7	0.9	**AUC**
Resnik	**sen**	0.6575	0.4714	0.3340	0.2602	0.1236	0.0747	0.0308	**0.684**
	fpr	0.4451	0.1111	0.0486	0.0301	0.0085	0.004	0.0005	
Lin	**sen**	0.9951	0.7381	0.5409	0.4198	0.2385	0.0979	0.0379	**0.7897**
	fpr	0.8627	0.2568	0.1214	0.0748	0.0314	0.0103	0.0032	
Rel	**sen**	0.7333	0.5195	0.3553	0.2656	0.112	0.0561	0.0202	**0.6889**
	fpr	0.5746	0.1396	0.0612	0.0369	0.0097	0.0039	0.0007	
JC	**sen**	0.9937	0.7405	0.5423	0.4219	0.2341	0.1013	0.0344	**0.7955**
	fpr	0.8428	0.2487	0.119	0.0737	0.0296	0.0101	0.0031	
Wang	**sen**	0.9025	0.6695	0.5156	0.4058	0.222	0.126	0.0601	**0.7339**
	fpr	0.846	0.2377	0.1197	0.0767	0.0268	0.0123	0.0035	
ShPath	**sen**	0.8166	0.628	0.4886	0.3997	0.2802	0.2044	0.1546	**0.8152**
	fpr	0.3736	0.0934	0.0412	0.0254	0.0077	0.0036	0.0005	
Cosine	**sen**	0.9246	0.6750	0.5118	0.4076	0.2148	0.1111	0.039	**0.7429**
	fpr	0.8455	0.241	0.1199	0.0778	0.026	0.0097	0.0027	

TABLE II: PROTEIN FUNCTION PREDICTION RESULTS BASED ON SEMANTIC HIERARCHICAL CLUSTERING WITH PIN

M	ω =	0	0.1	0.2	0.3	0.5	0.7	0.9	**AUC**
Resnik	**sen**	0.9996	0.7177	0.5351	0.4182	0.2114	0.0969	0.0209	**0.7549**
	fpr	0.9840	0.2799	0.139	0.0899	0.03	0.0111	0.0029	
Lin	**sen**	0.9996	0.7177	0.5351	0.4182	0.2114	0.0969	0.0209	**0.7549**
	fpr	0.984	0.2799	0.139	0.0899	0.03	0.0111	0.0029	
Rel	**sen**	0.9996	0.7177	0.5351	0.4182	0.2114	0.0969	0.0209	**0.7549**
	fpr	0.984	0.2799	0.139	0.0899	0.03	0.0111	0.0029	
JC	**sen**	0.9996	0.7177	0.5351	0.4182	0.2114	0.0969	0.0209	**0.7549**
	fpr	0.984	0.2799	0.139	0.0899	0.03	0.0111	0.0029	
Wang	**sen**	0.9851	0.7106	0.5322	0.4122	0.224	0.1197	0.0456	**0.7432**
	fpr	0.9954	0.2787	0.1464	0.0888	0.0355	0.0165	0.005	
ShPath	**sen**	0.9996	0.7177	0.5351	0.4182	0.2114	0.0969	0.0209	**0.7549**
	fpr	0.984	0.2799	0.139	0.0899	0.03	0.0111	0.0029	
Min	**sen**	0.9996	0.7177	0.5351	0.4182	0.2114	0.0969	0.0209	**0.7549**
	fpr	0.984	0.2799	0.139	0.0899	0.03	0.0111	0.0029	

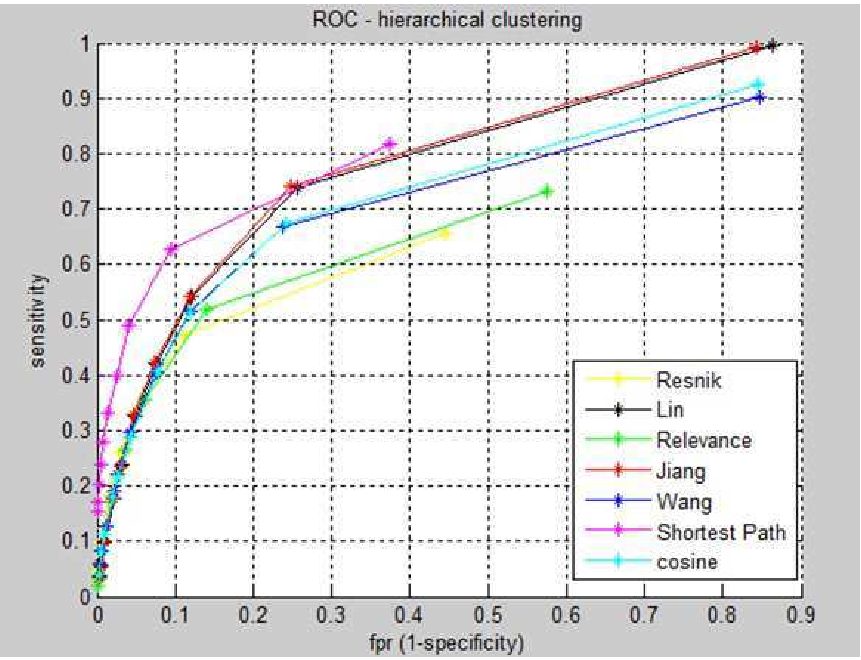


Fig. 2. ROC curves for evaluation of annotation using semantic hierarchical clustering without PIN

Figure 2 shows the semantic similarity metrics comparison ROC curves for evaluation of annotation using semantic hierarchical clustering without PIN.

Table 2 shows protein function prediction results for sensitivity and false positive rate, based on semantic hierarchical clustering with protein interaction network. It is evident from the table 2 that in this case only semantic similarity metrics don't give an improvement in the AUC value, which indicates that with graph clustering, semantic similarity metrics do not affect the protein function prediction process. Also, this shows that semantic similarity metrics have a bigger impact on protein function prediction based on hierarchical clustering of the dataset without the use of PIN, compared to protein graph hierarchical clustering with the use of PIN.

V. CONCLUSION

A new method for protein function prediction using semantic similarity metrics with and without protein interaction networks was presented. It is based on a clustering algorithm (hierarchical clustering) using a semantic similarity metric as a weight factor in the distance clustering matrix. The influence of 6 different semantic similarity metrics was evaluated with the protein data. Experiments revealed that the prediction accuracy of the method based on protein clustering without PIN with the use of hybrid semantic similarity metrics outperforms the use of other standard non-semantic similarity metrics. However, with PIN data and graph clustering, results show that semantic similarity metrics do not influence overall protein function prediction, but only give a small decrease in the false positive rate. High sensitivity gives a high false positive rate, therefore, the decrease of the false positive rate is a significant result. These results show the future need of detecting true protein interactions in PINs with the help of other semantic similarity metrics.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the “Ss. Cyril and Methodius” University, Macedonia.

REFERENCES

[1] The Gene Ontology Consortium: http://www.geneontology.org/U, Gene Ontology Documentation [Accessed March 2014].

[2] Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcão, A.O., Couto, F.M. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 2008, 9(Suppl 5): S4, 2008.

[3] Pesquita, C. Improving Semantic Similarity for Proteins based on the Gene Ontology. Master Thesis, University of Lisbon, Portugal, 2007.

[4] Brohée, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics 7*:48 (2006).

[5] Lin, C., Cho, Y., Hwang, W., Pei, P., Zhang, A.. Clustering methods in protein-protein interaction network. *Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*, John Wiley, Sons, Inc., Chichester, 2006.

[6] Resnik, P. Using information content to evaluate semantic similarity. *Proceedings of the IJCAI05*, pg. 448 – 453, 1995.

[7] Lin, D. An information-theoretic definition of similarity. *Proceedings of the 15th Int. Conf. on Machine Learning*, 1998.

[8] Jiang, J. Conrath, D.W. Semantic Similarity based on corpus and lexical taxonomy. *Proc. Of 10th Int. Conf. COLING*, 1997.

[9] Schlicker A., Domingues F., Rahnenfuhrer J., Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006, 7:302, 2006.

[10] Wang J.Z., Du Z., Payattakool R., Yu P.S., Chen C.F. A new method to measure the semantic similarity of GO term. *Bioinformatics* 2007, 23(10): 1274-1281, 2007.

[11] Shen Y., Zhang S., Wong H.S., Zhang L. A new method for measuring the semantic similarity on Gene Ontology. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2010,* pp. 533–538, 2010.

[12] Trivodaliev, K., Chingovska, I., Kalajdziski, S., Davcev, D. Protein Function Prediction Based on Neighborhood Profiles. *Proceedings of the ICT Innovations*, Springer-Verlaang, Macedonia, 2009.

Self-sustainable robotic platform powered by solar energy

Vladimir Popovski
Student at Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, Macedonia
vladimirpopovski@live.com

Nevena Ackovska
Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, Macedonia
nevena.ackovska@finki.ukim.mk

***Abstract*—This paper presents the process of creating a mobile robotic vehicle powered by solar energy. It also elaborates the process of using solar panels to supply the controller with energy and presents some intelligent decisions using its sensors and actuators. The platform is made of Lego Technic Bricks and uses Arduino controller connected with different sensors and actuators. The paper explains the sustainable usage of the robotic system in real environment. This paper also gives some insight on how this system can be improved, how different algorithms can be applied and how this robot can be used in the environment.**

Keywords*—*self-sustainable; robotic vehicle; solar panels; sun tracking;

I. INTRODUCTION

In the past two decades, many self-sustainable robots have been built mainly to execute various jobs that humans are unable to do. The first such robots are the rovers that were sent on Mars: Sojourner (1997) [1], Mars Exploration Rover - MER (2004), Spirit and Opportunity (2008), Mars Science Laboratory – MSL (2009) [2] and Curiosity (2012). Other self-sustainable robotic systems are used for exploration of areas that are uncommon to humans: LORAX (2005) [3], used for exploring microbes in Antarctica, and Zoë (2005) used for discovering life in Atacama Desert [4]. All of these robots are highly efficient, intelligent and precisely made to do some particular job. Many such solutions are built in smaller scales, in laboratory and home environments. One of these efforts is described in this paper. This paper presents a solution for creating a self-sustainable robot that can easily be modified during each stage of development. The work presented here is a follow up on our previous work [5].

The benefit of using renewable energies is the fact that they can be used again and again, with little consideration for the reserves. In other words, we have infinite amount of energy to harvest. According to a study given in [6], solar energy is the greatest energy source giving a total (including the loss in the clouds) of 23,000 TWy/year which is 25 times more than the total reserves of coal (900 TWy). If it is assumed that the robots are indestructible, those who use the solar energy (or other renewable energy) as their main energy source can live and operate eternally. This is the starting point in this research, which represents an experiment in creating robotic platform that can use the solar energy as a driving force.

The paper is organized in several sections. First, the physical construction of the body is discussed, describing the LEGO bricks as solution for building robots with body that can be easily changed by adding or removing LEGO structures. In this part also all the other sensors and actuators are given. After that the modeling of intelligent behaviour is presented. The possible further modifications are discussed in the future work section. The last part concludes the paper.

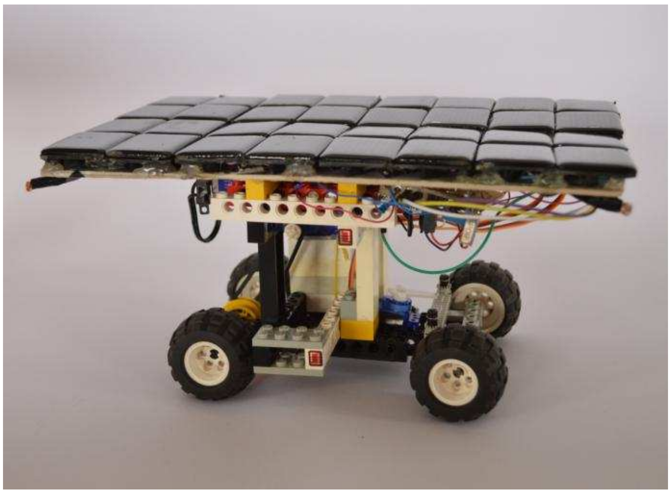


Fig. 1. The updated robotic system made of Lego bricks, Arduino controller, various circuits, sensors and actuators

II. BUILDING THE ROBOTIC PLATFORM

In our previous work, a robotic platform has been presented which consists of LEGO bricks as the main rigid body capable of upgrading. PV (photovoltaic) solar panels were the main energy supply for the 5V electromotor which move the robotic system. Since then the robot has been upgraded. The complete upgraded robotics platform is presented in Fig. 1. The robot was improved by adding an Arduino controller, two new electromotors used to move the system in all planar directions, an integrated chip that manage the use of different voltages and four photo-resistors used for modeling the intelligence. In the following sections all of these parts (Fig. 2) are presented in details.

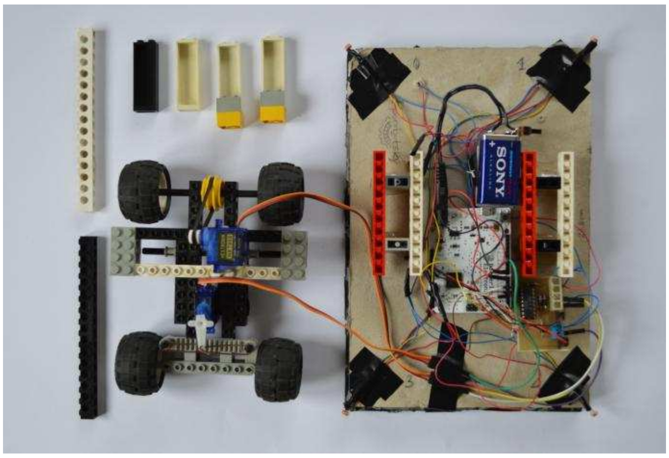


Fig. 2. Parts of the robotic system

For energy supply of the system that is renewable, solar energy is used, harvested by photovoltaic panels. 32 small (3cm X 4cm) solar panels are used producing 2 volts and 0.04 amperes and giving a total of 0.08 watts. These panels are grouped in two groups connected in series, each containing 16 small panels connected in parallel. In this way, the total of 4 volts and 0.64 amperes are produced, giving 2.56 watts.

The brain of each robot is its main controller. For this robot the Arduino Leonardo controller is used. Arduino Leonardo controller has ATmega32u4 processor with memory of 32 KB. It operates on 5 volts and the recommended input voltage is 7 – 12V. This controller has 6 input analog channels and 14 input/output digital channels. Seven of the digital channels can be used for PWM (pulse-width modulation). Since this controller required 7 – 12 volts to operate and the produced voltage by the solar panel are approximately 4V, an additional 9V battery is added on the system to supply the controller. This implies the fact that at this stage of development the system is not completely autonomous since it has non-rechargeable battery powering the controller. However, the main consumers of energy in the system are the motors, and they use the renewable energy from the panels. Since the controller does not require lot of energy, this battery will last a long time and it will determine the lifetime of the robot until it is replaced.

The motion of the robot is provided by using two servo electromotors HEXTRONIK HXT900 9GR. One of them is used for moving the vehicle forward and backward and the other is used to control the directions left and right. The first motor has been modified since these motors have limited rotation by 180 degrees. By removing its embedded potentiometer, a plane, yet powerful electromotor was created that can move in both directions 360 degrees. The second motor was attached to a mechanism connected with the front wheels, and its job is to move the wheels by receiving commands from 0 to 180 degrees. The motors are the greatest energy consumers in this robotic system. By testing the system it is concluded that the servo motor, that moves the wheels left and right, has an easy job and does not requires lot of energy. On the other hand, the main electromotor that moves the system forward and backward requires greater amount of energy to move the system.

The motors can be attached directly to the controller and drag energy of it. However, the controller is not capable of providing the required power for their movement. The decision was to separate the energy supply of the controller and the motors. This enables possibility of using different energy sources if needed, and makes the robot robust. For example, at night using solar energy is not an option, thus one can use different energy types for operating motors or controller if needed.

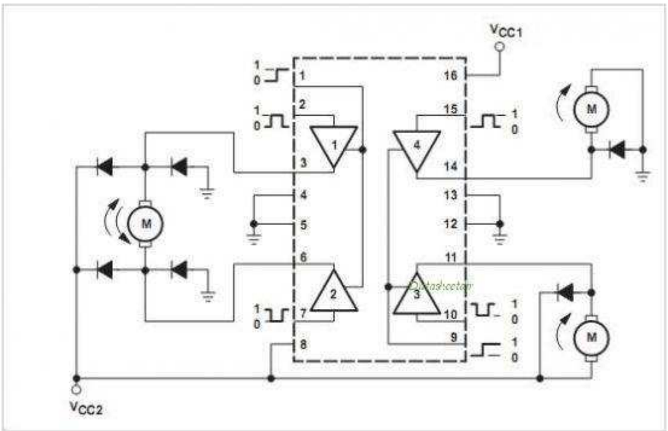


Fig. 3. L293N chip along with the protective diodes and electromotors

In order to provide a voltage from the panels but still control from the Arduino board, the L293-N integrated chip (fig. 3) is used. As shown in figure 3, L293-N contains four controllable modules for controlling four electromotors in one direction, two electromotors in both directions left and right or any combinations of the previous two modules. For this robotic system two modules are used for controlling the main electromotor in both left and right directions and one more module to control the servo motor. Along with the integrated chip, two diodes are needed to protect the chip from the induced current that is produced with the rotations of the motor. Pin 1 and 9 of the integrated chip are used as an on/off switch for the two motors. Pins 2, 7 and 15 are used for powering up the motors. The servo motor requires additional information from the controller about the degrees for rotation which is enabled by pin 11 of the Arduino controller. The complete abstracted circuit schematics can be seen on figure 4.

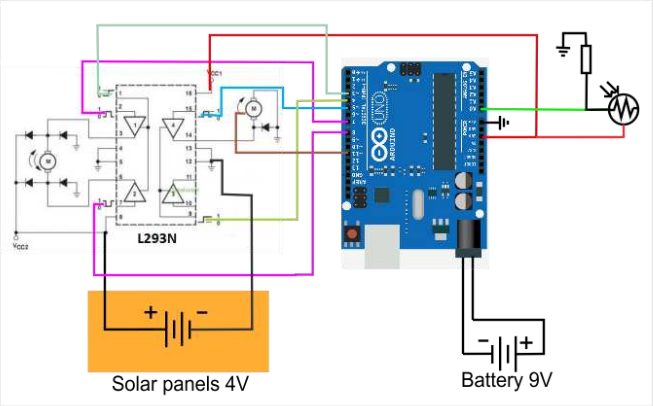


Fig. 4. Abstracted schematics of the system

At the end, the light sensors are added to the robot, each set on a different side of it. They are directly connected to the analog inputs of the controller providing information about the intensity of the light in the area. Here, additional resistor is used, as shown in figure 4, to regulate the input information to the system.

III. Modeling intelligence

Even nowadays defining the term intelligence is not simple. The intelligence involves many interconnected aspects like judgment, common sense, initiative, adoptability, emotions etc. When an object is observed it can be said that the type and the number of the sensors play a big role in its intelligence. The Alien view theory says “If you live in a world where the main sensory input is infrared signal, you may consider humans as non-intelligent beings” [7].

Having only four sensors, our goal was to model the intelligent behaviour by allowing the system to follow the most intense light in four directions. The sensors are placed on the four edges of the PV solar panels.

The main action of the robot is to move forward or backward. The system decides to move forward or backward depending on which sensors show bigger value of light received. Secondary movement is achieved by moving left and right. This movement only corrects the main action until the values of the both front or both back sensors show approximately same values (Fig. 5). After that is achieved the robot continues to move forward or backward until some different input from the sensors is received. If all of the sensors give approximately same values, the system will stop, showing that it is currently in the best surviving position.

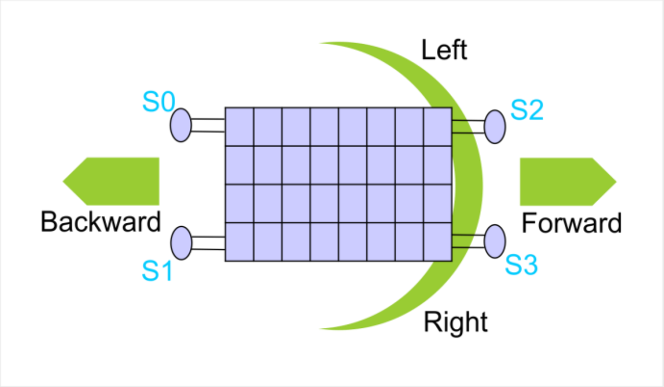


Fig. 5. Movement of the system

The Arduino controller is easily programmable environment, using C programming language. For the servo motor there is a servo library which can be imported into the code, allowing the use of many useful functions that can move the servo motor. By directly writing the degree (from 0 to 180 degrees) the servo motor moves the wheels left and right or straighten them so the robot can move in all directions.

The light sensors (photo resistors) provide information about the light received with number from 0 to 1024. Every two seconds the values of the sensors are updated and the controller decides which action to take next depending on the input values. The values can be modified by changing the value or the resistor connected from the negative terminal of the sensor to the ground.

IV. Testing the system

After assembling the system and packing its component in its compact body, some tests were provided to investigate the capabilities of the constructed robot. The system was tested on two different ways: outdoor and indoor.

The system was placed in an open-spaced area on a sunny day and it was observed. The robot was showing intelligent behaviour by following the side of the sun. To test its ability to follow the light more precisely, the system was placed indoor and an additional light source was brought and a movement of the sun was simulated. The system successfully followed the light.

Each of the motors require particular amount of energy. If the sun/light does not hit the panels directly, the voltage falls and the energy supply cannot support the system. This is taken into consideration for the next improvements steps where additional rechargeable battery will also support the motors.

V. Further improvements

The intelligence of a robotic system is connected with the number and types of sensors added to the system. First goal of improving this robotic vehicle is making it more intelligent by adding more sensors, flash memory and computational logic. This would help the robot better to perceive the surrounding environment and make more intelligent decisions. With more sensors also the logic of the robot will be more complex, making the system capable of dealing with various problems.

The panels are the most productive when the angle between the sun’s rays hitting the panels are 90 degrees. In our future work it is planned a rotating panels to be placed which will rotate and follow the sun.

Our robotic system can operate only when there is clear sun. Next thing to do is adding a battery as additional supporting power source to ensure the operability of the robot during night time or on a cloudy day. This way, the robot can operate during the day and also fill the battery, and when there is not enough sun it can rely on its additional power supply. When the battery is empty the system should stay in one position until the sun comes up again.

VI. Conclusion

This paper presents a robotic system that was built from Lego bricks, Arduino microcontroller, four sensors and two actuators. The goal was to model intelligent behaviour according to which the robot moves, using only small amount of sensors. The goal was reached by observing the positive results that were received from the tests.

The effort presented here is important since the robots that operate on renewable energies have the ability to sustain and operate without human help in special environmental conditions. This imposes the application of working in dangerous environments such as destroyed nuclear stations, as well as working in areas where living resources are low like deserts, polar regions, volcanoes and outer space. In such environments self-sustainable intelligent robots can explore, research, help victims and do many other useful things.

The robot presented in this paper is not yet capable of operating under extreme conditions and in rough terrains, but it gives the idea of building such robot in home-made laboratory. It is proof that not only the world dominant robotic laboratories, but almost everyone can contribute in the process of building brighter future by creation self-sustainable systems that help humanity.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, R. Macedonia.

REFERENCES

[1] Brian Wilcox, Tam Nguyen "Sojourner on Mars and Lessins Learned for Future Planetary Rovers", 28th Int'l Conference on Environmental Systems, Danvers, MA, July 1998.

[2] Max Bajracharya, Mark W. Maimone, Daniel Helmick "Autonomy for Mars Rovers: Past, Present and Future", Jet Propulsion Laboratory, California Institute of Technology, 2008.

[3] Liam Pedersen, David Wettergreen, Dimitrios Apostolopoulos, Chris McKay, Matthew DiGoia, "Rover Design for Polar Astrobiological Exploration", Robotics Institute, Carnegie Mellon University, 2005.

[4] David Wettergreen, Michael Wagner, Dominic Jonak, Vijayakumar Baskaran, Matthew Deans, "Long-Distance Autonomous Survey and Mapping in the Robotic Investigation of Life in the Atacama Desert", Robotics Institute, Carnegie Mellon University, 2005.

[5] Vladimir Popovski, Nevena Ackovska, "A Robotic System Povered by Solar Energy", The 10th Conference for Informatics and Information Technology 2013.

[6] Richard Perez, Mark Perez, "A Fundamental Look at Energy Reserves for the Planet", John Brian Shannon website, 2009.

[7] Kevin Warwick, "Atificial Intelligence, the Basics", Routledge 2012

Short-term electricity load forecasting of the Macedonian electric power system using neural networks

Aleksandra Dedinec
Institute of Intelligent Systems
Faculty of Computer Science and Engineering
Skopje, Macedonia
aleksandra.kanevche@finki.ukim.mk

Slobodan Kalajdziski
Institute of Intelligent Systems
Faculty of Computer Science and Engineering
Skopje, Macedonia
slobodan.kalajdziski@finki.ukim.mk

***Abstract*—Electric power companies can make best decisions in terms of unit commitment, planning of the generation and the maintenance of the system using electricity load forecasting. Also, it is of great importance that power companies have knowledge of the future demand with great accuracy in order to optimize the operation, minimize the financial risk and maximize the reliability of the system. Some algorithms for data mining play an important role in predicting energy consumption. In this paper the application of artificial neural networks as a tool for short-term electricity load forecasting is analyzed.**

The model is applied to the data for the Macedonian hourly electricity consumption for a period of five years (from 2008 to 2012). The data is divided into two groups of consumption: direct consumers (directly connected to the transmission network) and consumers connected to the distribution network. In this paper, a comparison of the error in the prediction of the electricity consumption of the first group with the error in the prediction of the consumption of the second group of consumers is made. Also, the results are compared with the corresponding predictions given by the Electricity Transmission System Operator of Macedonia (MEPSO).

Keywords*— *artificial neural networks; electricity load forecasting; short-term forecasting

I. INTRODUCTION

Short-term load forecasting refers to minute, hour, day or week ahead predictions of the electric load. The forecasting of the load greatly improves the real-time control and the security of the system, as well as the efficiency of the system. Also, by having accurate prediction of the electric load, the price of the electric power can be significantly reduced. Especially, now after the liberalization of the electric power industry, these predictions play important role in decision making for both the power system operators and the market participants [1].

Artificial neural networks are widely used for load forecasting. Generally, they do not require detailed information about the system. Instead they employ process of learning, by which the relationship between the input and the output variables is defined. During the learning phase, this complex input-output mapping does not require detailed programming and defining the linear or nonlinear relationships between the data sets [2]. They are solely able to ignore the information that has little significance to the output and to focus on the variables that have high influence on the output [3]. Furthermore, the neural networks are highly parallel, so the results are obtained at high speed [2].

In this paper, the model of load forecasting using neural networks is applied to the data of the Macedonian hourly electricity consumption for a period of 5 years (2008-2012). The consumption data is divided into two groups. The first group is represented by the consumers that are directly connected to the electric power transmission network, which mainly involves the big industry companies. The second group represents the consumers that are connected to the distribution network, where one of the main consumers is the household sector. In this paper, it is analyzed which of these two groups is more predictable. The Electricity Transmission System Operator of Macedonia (MEPSO) provides information about their prediction of the electricity load for both groups [4]. So, the results of the neural network forecasts are compared to the forecasted values of the Macedonian system operator (MEPSO) and to the actual data.

Short-term load forecasting using neural networks attracts much attention in the literature. In [5] the study of the part of the Turkish power system is analyzed. In [3] the Taiwan power system load is predicted and part of the Italian power system load is forecasted in [6] using neural networks. The model of neural networks is compared to other models as in [7], and modifications of the classical neural networks model is widely analyzed in the literature as it is in [8-11]

The structure of the paper is the following. In the next section the application of the neural networks to electricity load forecasting is presented. In the third section the results and corresponding discussion are presented. The last chapter concludes the paper, and some plans for future work are described.

II. NEURAL NETWORKS

The most popular form of NN is the so called multi-layer feed-forward perceptron (MLP) structure, which means that the network is represented as a directed acyclic graph, whose structure has several layers - an input layer, one or several hidden layers and an output layer. The input layer gathers the model's inputs vector **x** while the output layer yields the model's output vector **y** **[8].** Usually there is only one hidden layer, because it is the best solution for the majority of the practical problems **[5].**

There are neurons in the hidden layer which are activated by a non-linear function. Most usually the tangent hyperbolic function is used. [8]

So, if there is a neural network with *d* input variables, *h* hidden layer neurons and a single linear output variable *y*, the non-linear mapping between the input ***x*** and the output *y* is given by the following equation:

$$y = \sum_{j=0}^{h}\left[w_j f\left(\sum_{i=0}^{d} w_{ji} x_i \right)\right]$$

The parameters w_j and w_{ji} represent the weights and biases that connect the layers [8].

The main advantage of the artificial neural networks is the ability to learn. For the purpose of forecasting and prediction which belong to the regression analyses, the supervised learning is used. In this case the cost function is function of the difference between the forecasted and the actual data. In the training (learning) phase the goal is to adjust the weights in order to minimize the cost. The most widely cost used is the mean-squared error, given by the following equation:

$$E_D = \frac{1}{2}\sum_{i=1}^{N}\{y_i - t_i\}^2 = \frac{1}{2}\sum_{i=1}^{N} e_i^2$$

Where y_i is the actual data and t_i is the forecasted data.

In the learning process the weights are updated in order to minimize the error, which requires minimizing a non-linear function. The numerical solution to this problem may be provided by different algorithms, as the steepest descent algorithm, Newton's method, Gauss–Newton's algorithm, and Levenberg–Marquardt algorithm. The Levenberg–Marquardt algorithm is a combination of the steepest descent method and the Gauss–Newton algorithm. Fortunately, the speed advantage is inherited from the Gauss–Newton algorithm and the stability from the steepest descent method. It's more robust than the Gauss–Newton algorithm. The Levenberg–Marquardt algorithm converges a little bit slower than the Gauss–Newton algorithm, but it converges much faster than the steepest descent method. Because of these advantages the Levenberg–Marquardt algorithm is used, in this paper, for updating the weights, as one of the most fastest and stable algorithms. [12]

The next phase is the generalization phase [8], which means testing the neural network whether it is able to generalize, which means whether it is able to give correct output for a certain input. The most commonly metric used in the electricity load forecasting are MAPE (mean absolute percent error) and MAE (mean absolute error), which are given by the following equations:

$$MAE = |e_i| = |y_i - t_i|$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|e_i|}{t_i} \times 100 \right)$$

So, the process of setting up the neural network for load forecasting reduces to the following steps [3]:

1. *Select input and define output variables.*

2. *Determine layer(s) and the number of neurons in hidden layers*. As previously mentioned the number of layers used in this paper is three. The best results for the number of neurons in the hidden layers are achieved with 20 neurons.

3. *Learning (or training) from historical data.* - the process of adjusting the weights in order to minimize the cost, i.e. the error.

4. *Testing (Globalization).* After the training phase, the neural networks are tested whether the general rule is learned, which means whether the neural network gives the desired output, given a certain input.

One of the major steps that need to be properly defined in order to obtain good results is to determine the input and output variables. When it comes to hourly load forecasting, there is only one output variable, which is the consumption of electricity in that hour. The determination of the input variables is not so straight forward, and should be extracted from the analyses of the historical data of electricity load.

A. Selection of input variables

Because there is no general rule that can be followed in order to define the input variables, same statistical analyses which variables have influence on the electricity load. For that purpose, we have analyzed the electricity load of the Macedonian electric power system.

The load is changing on a daily basis. Each day the load is increasing during the day, and decreasing during the night. There is a peak load, which typically occurs in the afternoon, when the people come back from work. An average daily load for the year of 2011 is presented on Fig. 1. Naturally one of the variables should be the hour indicator. Also, the load of the same hour the previous day should give significant information.

The load is also changing during one week. The load during the weekends is much less than the load during the working days. So, very important indicator is the day of week variable. Also, the previous week same hour load is very significant. On Fig. 2 the total daily load of a typical week in 2012 is presented. It can be noticed that on Tuesday the load is drastically reduced. This is because that day is a holiday in Macedonia (1-st of May). So, the information whether it is a holiday or not drastically changes the load, and it should be also one of the input variables.

The load is also changing depending on the season, which is the period of the year. This means that the daily load is changing slowly during the year. Because of this the previous 24 hours average load is also considered as one of the variables.

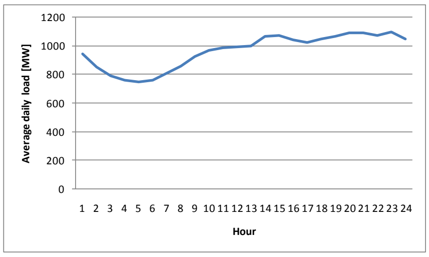


Fig. 1 Average daily electricity load for the Macedonian electric power system for 2011

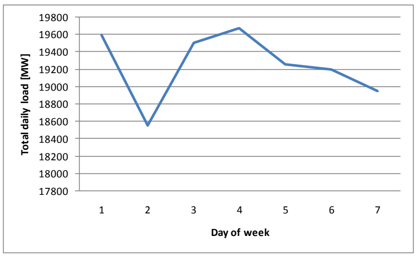


Fig. 2 Total daily load of a typical week in 2012

Based on the previous analyses of the data and the literature review [2,3,5,13], on Fig. 3 the final structure of the neural network, used in this paper, is presented.

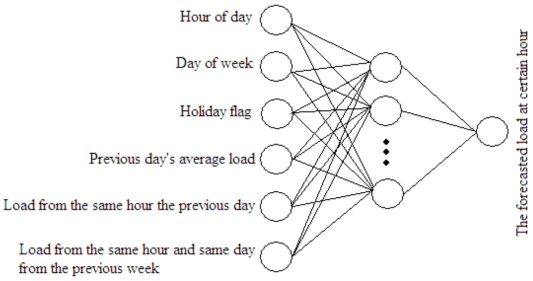

Hour of day
Day of week
Holiday flag
Previous day's average load
Load from the same hour the previous day
Load from the same hour and same day from the previous week
The forecasted load at certain hour

Fig. 3 Multilayer neural network for load forecasting

III. Results

For modeling the load forecasting using neural networks MATLAB software is used [13,14]. The training period includes the electricity load during the period of 2008-2011. The testing of the neural network results is done for the first three months of 2012.

Fig. 4 represents the results of the total hourly load forecasting using the model of neural networks which was described in the previous section (yellow line). Also, the reported predictions about the electricity load by the Macedonian system operator (MEPSO) are presented by red line. The actual data for the load are presented by the blue line. The absolute error of the prediction of the load using neural networks, which is the difference between the actual and the NN forecasted load, is presented on Fig. 5 by blue line. The absolute error of the load prediction using the MEPSO data is presented by the red line. It is obvious that the error is smaller when the neural network model is used.

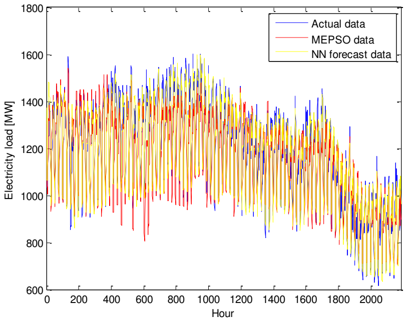


Fig. 4 NN forecasted data, MEPSO data and the actual data for the electricity load during the test period

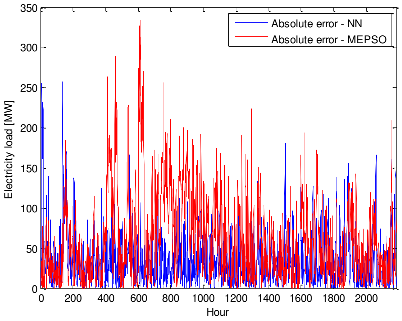


Fig. 5 The absolute error of the neural network and MEPSO forecasting

Quantitatively, the results are compared using the metrics described in the previous chapter, i.e. MAE, MAPE and daily peak MAPE. The results for the direct consumers, the distribution network consumers and the total load are presented on Table 1. It can be noticed, that the consumption of the direct consumers connected to the transmission network is not well predicted by the neural network. Better results are presented by the system operator. On the other hand, the distribution network consumers, which mainly represents the household sector is better predicted by the neural network model than by the system operator. Because bigger part of the total load is the distribution network consumers, the results of the neural network predictions are better if neural networks are used.

Table 1 MAE, MAPE and Daily peak MAPE

	Direct consumers		*Distribution network consumers*		*Total load*	
	NN	*MEPSO*	*NN*	*MEPSO*	*NN*	*MEPSO*
MAE [MW]	15.00	15.15	35.00	59.08	34.92	61.25
MAPE [%]	9.34	9.31	3.51	5.88	3.04	5.20
Daily peak MAPE [%]	8.26	10.56	3.49	5.27	2.60	4.84

The error distribution, the absolute error distribution and the absolute percent error distribution of the electricity load forecasting using neural networks is presented on Fig. 6.

It is interesting to investigate in which hour of the day the average error is the highest. On Fig. 7 the average load error depending on the hour of the day is presented. It can be noticed that the highest error in the predictions is obtained in the afternoon between 13:00 and 18:00 for the two analyzed groups of consumers as well as for the total consumption.

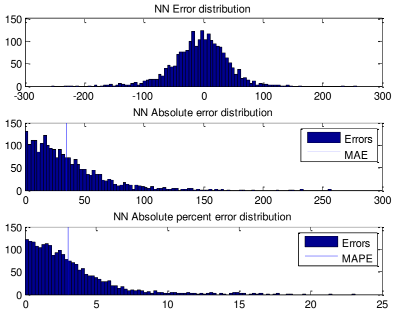


Fig. 6 Error distribution, absolute error distribution and absolute percent error distribution using neural network

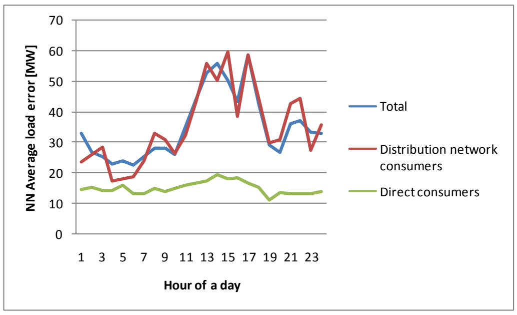


Fig. 7 Average load error depending on the hour of the day using neural networks

The average load error depending on the day of the week is presented on Fig. 8. For the households the most unpredictable load is during the Saturdays, as well as Mondays. For the big industry consumers the highest error in the prediction is obtained on Monday.

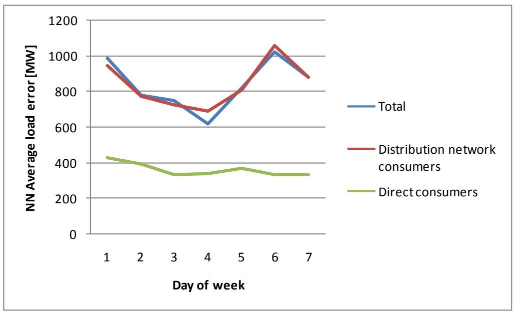


Fig. 8 Average load error depending on the day of the week using neural networks

IV. CONCLUSION AND FUTURE WORK

The results showed that the neural network models obtain good prediction of the electricity load. The results for the total hourly load are better compared to the predicted values of the Macedonian system operator. The big industry companies are more unpredictable by using neural networks, than the household consumers.

In the future, the results may be improved by implementing more input variables. One of them may be the weather information, i.e. the outside temperature, which influences the electric power consumption for heating and cooling. Also, because the household is the biggest consumption sector and it is highly influenced by the information whether the electric power price is according to cheap or expensive tariff, this information may also be included as an input variable. As well as for load forecasting, this model may be applied for electricity price forecasting, which may be of much interest after the total liberalization of the Macedonian electric power market.

V. ACKNOWLEDGEMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University.

REFERENCES

[1] A. Deihimi, O. Orang, and H. Showkati, “Short-term electric load and temperature forecasting using wavelet echo state networks with neural reconstruction,” Energy, vol. 57, pp. 382-401, August 2013.

[2] N. Kandil, R. Wamkeue, M. Saad, and S. Georgeas, “An efficient approach for short term load forecasting using artificial neural networks,” International Journal of Electrical Power & Energy Systems, vol. 28, No. 28, pp. 525-530, October 2006.

[3] C. Hsu, and C. Chen, “Regional load forecasting in Taiwan-- applications of artificial neural networks,” Energy Conversion and Management, vol. 44, No. 12, pp. 1941-1949, July 2003.

[4] Electricity Transmission System Operator of Macedonia, URL: http://www.mepso.com.mk/

[5] T. Yalcinoz, and U. Eminoglu, “Short term and medium term power distribution load forecasting by neural networks,” Enegy Converion and Management, vol. 46, No. 9-10, pp. 1393-1405, June 2005.

[6] M. Beccali, M. Cellura, V. Lo Brano, and A. Marvuglia, “Forecasting daily urban electric load profiles using artificial neural networks,” Energy Conversion and Management, vol. 45, No. 18-19, pp. 2879-2900, November 2004.

[7] A. Badri, Z. Ameli, A.Motie Birjandi, “Application of Artificial Neural Networks and Fuzzy logic Methods for Short Term Load Forecasting,” Energy Procedia, vol.14, pp. 1883-1888, 2012.

[8] P. Lauret, E. Fock, R. Randrianarivony, and J. Manicom-Ramsamy, “Bayesian neural network approach to short time load forecasting,” Energy Conversion and Management, vol. 49, No. 5, pp. 1156-1166, May 2008.

[9] H. Su, “Chaos quantum-behaved particle swarm optimization based neural networks for short-term load forecasting,” Procedia Engineering, vol.15, pp. 199-203, 2011

[10] N. Tang, and D. Zhang, “Application of a Load Forecasting Model Based on Improved Grey Neural Network in the Smart Grid,” Procedia Engineering, vol. 12, pp. 180-184, 2011

[11] N. Kourentzes, D. Barrow, and S. Crone, “Neural network ensemble operators for time series forecasting,” Expert Systems with Applications, vol. 41, No. 9, pp. 4235-4244, July 2014.

[12] H. Yu, and B. Wilamowski, “Levenberg–Marquardt Training”, URL: http://www.eng.auburn.edu/~wilambm/pap/2011/K10149_C012.pdf

[13] A. Deoras, “Electricity Load and Price Forecasting Webinar Case Study”, 2011, URL: http://www.mathworks.com/matlabcentral/fileexchange/28684-electricity-load-and-price-forecasting-webinar-case-study/content/Electricity%20Load%20&%20Price%20Forecasting/Load/html/LoadScriptNN.html

[14] Matlab Neural Network Toolbox, URL: http://www.mathworks.com/products/neural-network/

Spatial Information System for Analysis of the Quality of Drinking Water in Republic of Macedonia

Elizabeta Kroneva
KomSoft dooel Computer Engineering -
Skopje, R. Macedonia

Kosta Mitreski
University" Ss. Cyril and Methodius "–
Faculty of Computer Science and Engineering
Skopje, R.Macedonia
kosta.mitreski@finki.ukim.mk

Abstract—**Environmental pollution (air, water, and soil) is a problem with which the whole world is preoccupied. The rapid rise of population, swift industrialization, transport, agriculture, and social inequality are the main reasons of the worsening of the quality of the environment. In this paper we will analyze the quality of drinking water in R. Macedonia. Monitoring of the water quality is done in places which represent a health interest, in order to undertake measures to prevent possible harmful effects on the health status of the population.**

In this paper will be represented a Spatial IS analysis of the quality of drinking water with data from analysis made from samples of water and display the results of the processing on a map of R. Macedonia.

Keywords—Quality of drinking water, Spatial IS, Quality management

I. INTRODUCTION

The use of visual modeling and analysis of the quality of drinking water with spatial information system is implemented in a large number of studies. In paper [1] is conducted an analysis of the quality of drinking water with the use of GIS in the city of Bhadravathi, India. The main supply of water for the population and the industry is from the river Bhadra. During the analysis 12 physical and chemical parameters were considered. Geographical information system in this study is used to represent the spatial distribution of these parameters. In study [2] is conducted an analysis of physical and chemical data of groundwater in a period before and after the monsoons. Samples were taken from 49 different wells. Spatial distribution of the maps was prepared for various physical and chemical parameters. In paper [3] is given an assessment groundwater quality with determining physical and chemical parameters (pH, EC, TDS, and TH) and major ion concentrations (HCO_3, Cl, FSO_4, Ca, Mg, Na и K) around Dindigul district, Tamil Nadu, India. Samples were taken from 59 wells covering the entire area. A GIS mapping technique was adopted to highlight the spatial distribution of physical and chemical parameters and major ion concentration in groundwater. In paper [4] is done testing and analysis of groundwater in Gulbarga, India using GIS as a tool that is used for storing, analyzing and displaying spatial data. For this study, 76 samples were collected from created wells and open wells. On the samples of water analyses were done for the physical and chemical parameters TDi, Th, CL- и NO3-. Maps with information on groundwater quality are prepared using GIS spatial interpolation technique for all the above mentioned parameters. The results obtained in this study and spatial databases made in GIS are useful for monitoring and management of groundwater pollution in the research area.

In this paper we use data collected from measurements of the Institute of Public Health (IoPH) of RM, will make data analysis for water quality using relational database represented by web-solution and appropriate analyzes and finally it will be modeled visually present with the tools used for this purpose (Google maps) and interoperability in Esri GIS for further analysis and visualization with this.

II. ARCHITECTURE OF THE SPATIAL IS

Fig. 1 illustrates a given UML diagram of the database. Also shown is the connection between tables by fields who in one table are primary keys. Each table that has a primary key has a field that is marked with {unique}.

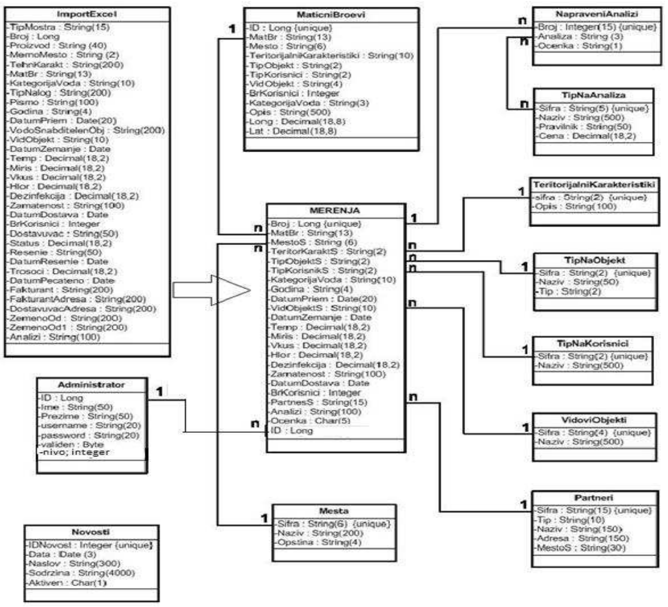


Fig. 1. UML diagram of the database

The place where the sample is taken of the water for the analyzes that are performed is called measuring point. Each measuring point has its own unique identification number.

Each measuring point registered with its own identification number has appropriate geographical coordinates (longitude and latitude) by which they are displayed on a map of R. Macedonia. Laboratory analysis that is made on the samples is entered via a web application. Web application consists of two: public and administrative

Every user or visitor to the web application has access to the public part. In this section an overview is provided of events news entered by the Institute for Public Health, biography and contact the professional team at the Institute of Public Health where the analysis are done. Administrative part of the section to which users have access only with a username and password and with that a higher level of access is given, in this case they are people who are employed at the Institute of Public Health. In this section it is possible to insert, change and review data. The user interface used to display the data is created using the HTML language and using CSS styles. The page is divided into five main sectors: DIV header, menu, content, right and footer used to display specific content.

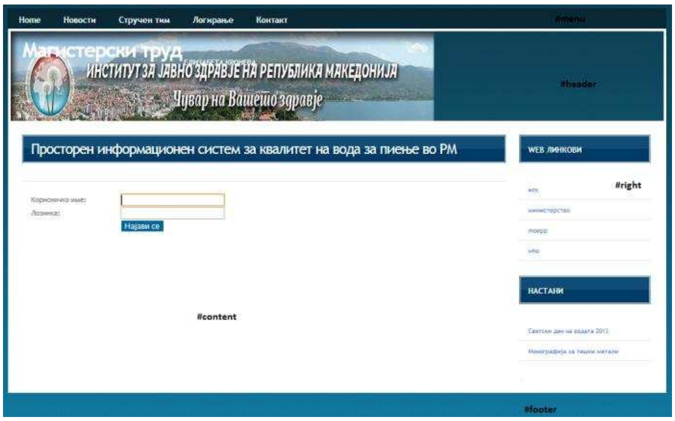
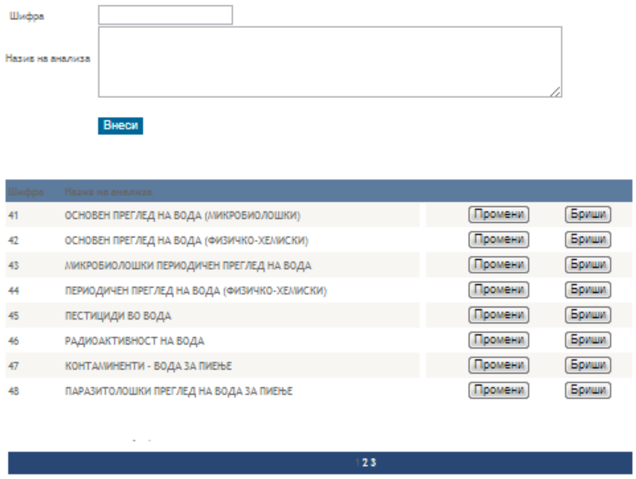


Fig. 2. User interface of the web-application

Analysis which affect the assessment of the quality of drinking water are microbiological, physical-chemical , pesticides in water, radioactivity on the water, contaminated drinking water, parasitological examination of drinking water. For each sampled can be done more analysis but even if a single one not satisfy MKD values, the final result of the analysis is assessed as negative and appropriate measures are taken.



Шифра	Назив на анализа		
41	ОСНОВЕН ПРЕГЛЕД НА ВОДА (МИКРОБИОЛОШКИ)	Промени	Бриши
42	ОСНОВЕН ПРЕГЛЕД НА ВОДА (ФИЗИЧКО-ХЕМИСКИ)	Промени	Бриши
43	МИКРОБИОЛОШКИ ПЕРИОДИЧЕН ПРЕГЛЕД НА ВОДА	Промени	Бриши
44	ПЕРИОДИЧЕН ПРЕГЛЕД НА ВОДА (ФИЗИЧКО-ХЕМИСКИ)	Промени	Бриши
45	ПЕСТИЦИДИ ВО ВОДА	Промени	Бриши
46	РАДИОАКТИВНОСТ НА ВОДА	Промени	Бриши
47	КОНТАМИНЕНТИ - ВОДА ЗА ПИЕЊЕ	Промени	Бриши
48	ПАРАЗИТОЛОШКИ ПРЕГЛЕД НА ВОДА ЗА ПИЕЊЕ	Промени	Бриши

Fig. 3. Interface for entering of Type of Analysis

All global entities are entered in the same way.

Excel file is placed on the server by the user with the highest level of access.

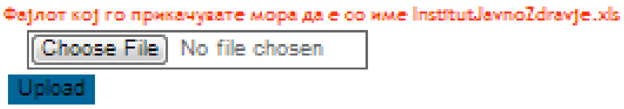
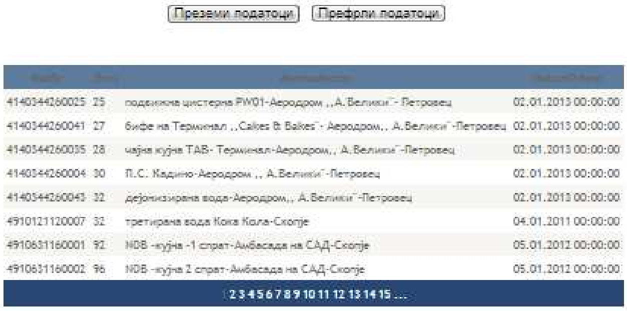

Фајлот кој го прикачувате мора да е со име InstitutJavnoZdravje.xls

Fig. 4. Appearance of the page for setting the file on the server

File which is placed on the server and contains data with performed analyzes have not yet been placed in the database. Through a special link on the website is enabled taking over data from the file and their presentation in the grid on the page as well as their transfer to the database.



4140344260025	25	подвижна цистерна PW01-Аеродром ,,А. Велики"- Петровец	02.01.2013 00:00:00
4140344260041	27	бифе на Терминал ,,Cakes & Bakes"- Аеродром,, А.Велики"-Петровец	02.01.2013 00:00:00
4140344260035	28	чајна кујна TAB- Терминал-Аеродром,, А.Велики"-Петровец	02.01.2013 00:00:00
4140344260004	30	П.С. Кадино-Аеродром ,, А.Велики"-Петровец	02.01.2013 00:00:00
4140344260043	32	дејонизирана вода-Аеродром,, А.Велики"-Петровец	02.01.2013 00:00:00
4910121120007	32	третирана вода Кока Кола-Скопје	04.01.2011 00:00:00
4910631160001	92	NDB -кујна -1 спрат-Амбасада на САД-Скопје	05.01.2012 00:00:00
4910631160002	96	NDB -кујна 2 спрат-Амбасада на САД-Скопје	05.01.2012 00:00:00

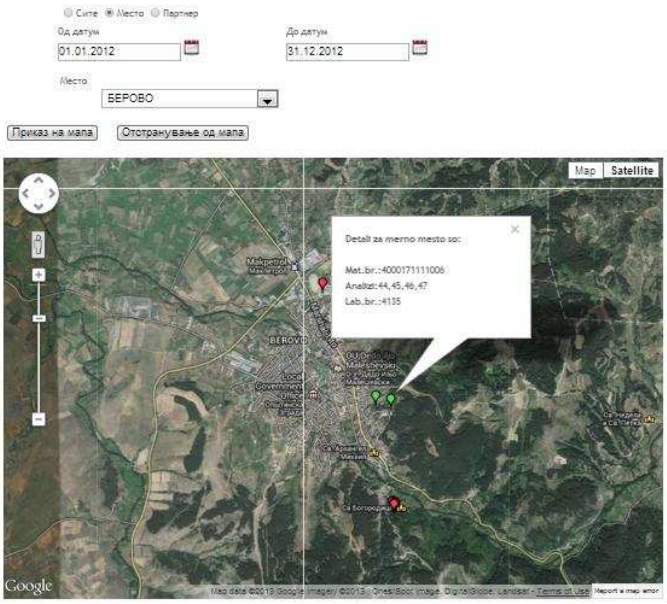
Fig. 5. Interface for transfer and download data in spatial IS

Presenting the results of the RM map by selected criteria is done in two ways using Google Maps and ArcGis Maps.

The user has three offered options by which will make a filter of assessments of the quality of the water for a certain period. Offered options are:

- Filtering only by selected period, for all positions in the Republic of Macedonia
- Filter by places for selected period
- Filter by partners for selected period

We will show an example of selected place Berovo for the whole period of 2012. By clicking the button "View" on the map markers will appear in the appropriate color red or green depending on the grade obtained. I must mention for the protection of the confidentiality of the results in the example data evaluation are not real.



○ Сите ● Место ○ Партнер

Од датум: 01.01.2012

До датум: 31.12.2012

Место: БЕРОВО

[Приказ на мапа] [Отстранување од мапа]

Detali za merno mesto so:

Mat.br.:4000171111006
Analizi:44,45,46,47
Lab.br.:4135

Fig. 6. Appearance of the page for presenting the results using Google Maps

Clicking on any of the markers in the field info we get information on which personal and laboratory number marker refers, as well as which analyzes are made by taking samples.

In case when more analyzes are made, but only one has been evaluated by derogation from the MKD final assessment of the analyzes for the sample is negative and is shown with a red marker.

The second way of presenting the results on the map of Macedonia is using the ArcGis. By creating Blank Page in ArcMap we can see that in Table of Content we have only Layers but further here we will add data and maps that we need. In Excel document we can see that there are more columns whose data for each column we would like to present on the map.

To enter data from Excel table in ArcMap first is necessary to enter the table that we want to work. Then on the places X Field and Y Field to select Long and Lat properly and in Z Field we chose the column whose data we want to display on the map. Then choose a coordinate system in which you work ie in our case the Geographic Coordinate Systems -> World -> WGS 1984th By confirmation on the above choices are displayed the data column that was selected in column Z Field (in this case, personal identification numbers) in the form on the points.

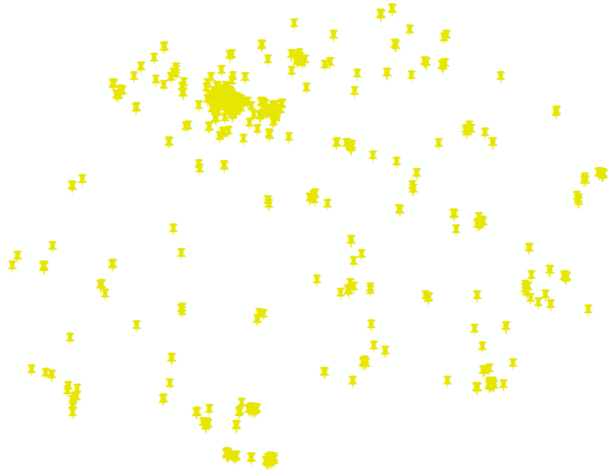


Fig. 7. Presentation of results for personal numbers

Next we have to do is to enter a map. In this case map Imagery is chosen which is shown on the Fig. 8.

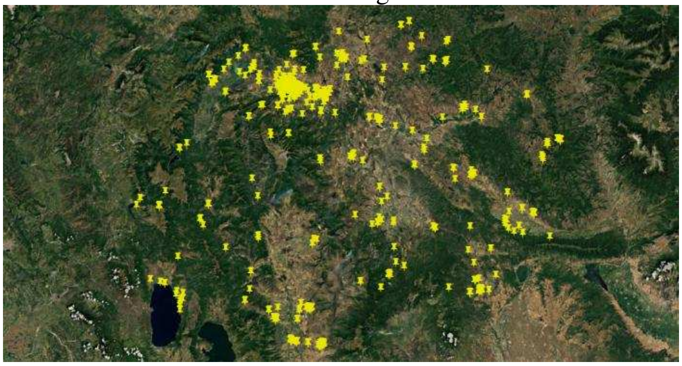


Fig. 8. Presentation of results for personal numbers using the map Imagery

After you have entered data in ArcGis, gets all the points from the table. But for example, it is necessary to show all the analyzes made during one year in a particular place, showing the positive analysis with green marker, negatively assessed analysis with red marker. As an example we will take the year 2012, the place Pehchevo. Selected data are exported. We use the same procedure for obtaining negative assessed analysis. We will use green color for positive assessed analyzes (Assessment = 0) and red color for negatively assessed analyzes (Assessment> 0).

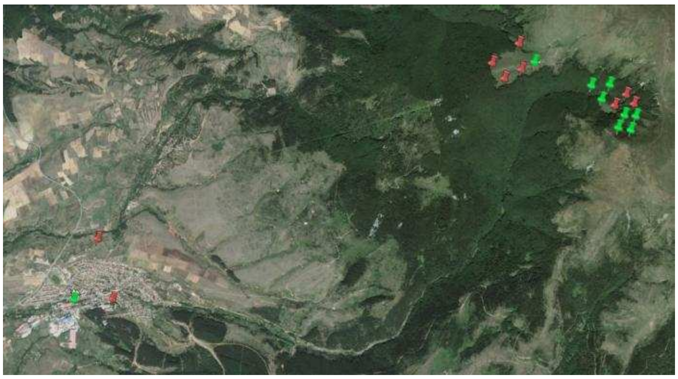


Fig. 9. Presentations of Results for Pehchevo 2012

III. CONCLUSION

By increasing the number of the technologies that are available, is increasing the development of the Web GIS applications. Although the technologies used for development of Web GIS are different, the power of geographic information system is the ability to visualize the real world. Allowing a large number of options and abilities for creating feedback with great accuracy and quality makes geographic information systems suitable for use.

In this paper Web application was presented for input data of performed analysis of the drinking water. Using GIS map provides a spatial overview of the analysis assessment for the selected period.

REFERENCES

[1] R. V. Raikar, M.K. Sneha, ”Water quality analysis of Bhadravathi taluk using GIS – a case study” May 2012. Available at http://www.ipublishing.co.in/ijesarticles/twelve/articles/voltwo/EIJES3230.pdf

[2] N. S. Magesh, N. Chandrasekar. Evaluation of spatial variations in groundwater quality by WQI and GIS technique: a case study of Virudunagar District, Tamil Nadu, India. Jun 2013. Available at http://link.springer.com/article/10.1007%2Fs12517-011-0496-z

[3] N.S.Magesh, S. Krishnakumar, N. Chandrasekar, John Prince Soundranayagam, Groundwater quality assessment using WQI and GIS techniques, Dindigul district, Tamil Nadu, India. November 2013. Available at http://link.springer.com/article/10.1007/s12517-012-0673-8

[4] P. Balakrishnan, Abdul Saleem, N. D. Mallikarjun, Groundwater quality mapping using geographic information system (GIS): A case study of Gulbarga City, Karnataka, India.(doi: 10.5897/AJEST11.134). December 2011.

[5] Nag.S.K, Poulomi Ghosh, Groundwater quality and its suitability to agriculture – GIS based casestudy of Chhatna block, Bankura district, West Bengal, India. Jun 2011. Available at http://www.academia.edu/1285019/Groundwater_quality_and_its_suitability_to_agricultureGIS_based_case_study_of_Chhatna_block_Bankura_district_West_Bengal_India.

Structure from motion obtained from low quality images in indoor environment

Bojan Dikovski, Petre Lameski, Eftim Zdravevski, Andrea Kulakov
Faculty of Computer Science
and Engineering
University Ss. Cyril and Methodius
Skopje, R. of Macedonia
Email: b.dikovski@yahoo.com, {petre.lameski, eftim.zdravevski, andrea.kulakov}@finki.ukim.mk

***Abstract*—Structure from motion is the process of extracting 3D structure from images taken through the motion of the camera. The result is dependent on the quality and resolution of the images that are being taken, so it is beneficial to use as high quality images as possible. Sometimes it is not possible to obtain high level of detail in photos because of the environmental, economic or other restrictions. In this paper we analyze structure from motion when using a low resolution camera in indoor environment. The obtained results are compared with the same process when using images of higher resolution and with the 3D structure created from points taken with the Microsoft Kinect sensor.**

Keywords - structure from motion, sparse reconstruction, low resolution images, Kinect.

I. Introduction

The process of extracting geometric structures from images taken through a camera motion has an extensive research history and already a few commercial systems are available [1][2]. The most notable work publsihed on this topic is the book *"Multiple View Geometry in Computer Vision"* [3], in chapters 9 through 12. Obtaining three dimensional structure from motion (SFM) is a similar problem with finding the structure from stereo (or multiview) vision. The difference is that in stereo vision we know what is the motion between the cameras, while in SFM this is not known. Calibrated stereo rigs in theory provide better and more accurate reconstruction, but SFM has an advantage in more simplistic recording procedure which is one of the main motivations to use it in this work.

Python Photogrammetry Toolbox and GUI (PPT) [4] is an open source software package that incorporates different tools needed to perform a SFM reconstruction. It includes the SIFT algorithm [5] to detect local features in the images which are then used by Bundler [6]. Bundler is a software that reconstructs the scene incrementally using a modified version of Sparse Bundle Adjustment [7]. The output of this is a sparse point cloud, but a cloud consisting of denser points can also be made using another software package called Patch-based Multi-view Stereo (PMVS2) [8]. PPT includes the PMVS2 package, as well as the Clustering Views for Multi-view Stereo (CMVS) software which can be used for preprocessing before PMVS2. Our intent is to explore the feasibility and performance of a SFM system with low to medium quality images of an indoor environment. Using free and open source software means that anyone could easily perform the reconstruction made in our experiment producing a 3D model of any normal household item while using images made with the camera of a low-end smartphone.

To measure the results from our experiments we chose to use the Microsoft Kinect [9] sensor for providing the ground truth data. It is a low cost device that comes with a RGB camera, depth and audio sensors. The accuracy of the Kinect depth data is comparable to a laser scanning data and does not contain large systematic errors which was shown through theoretical and experimental accuracy analysis in [10]. In this work it was concluded that for mapping applications the data should be acquired in the range of 1-3 meters distance from the sensor. This range is optimal for our experiment and the setup that we used. In [11] quantitative comparison of the Kinect accuracy with stereo reconstruction from SLR cameras and a 3D-TOF camera is done. The authors have concluded that the Kinect sensor performed close to, and in some cases overperformed, the other types of reconstruction.

The rest of this paper is organised in the following way: in section 2. we talk about related work in this research area, in section 3. we explain our experiment in great detail, in section 4. we show the results of the experiment, and finally in section 5. we give a conclusion.

II. Related Work

Structure from motion in the past has been applied mostly for estimation and remaking of three dimensional structures from images made in outdoor enviroments, usually at large distances from the target structure. [12] presents an approach for modeling and rendering existing architectural scenes from sparse sets of still photographs. This approach combines an interactive photogrametric modeling method and a model-based stereo algorithm which can create realistic views of architectural scenes even further away from the original photographs. In [13] a method with *O(n)* complexity has been proposed for organizing an unordered image set into clusters of related images from the same scene. The process of clustering is based on finding matches between the image features, similar to what is done in the process of structure from motion estimation in this paper. [14] introduces an automated large-scale image registration system that was used to create a large image dataset of the MIT campus. This data has been used in research for image-based rendering and 3D reconstruction. In [15] a framework based on a Bayesian model is used for automatic acquisition of three dimensional architectural models from short image sequences. Here an object recognition approach learns of the objects identity which can then be used to

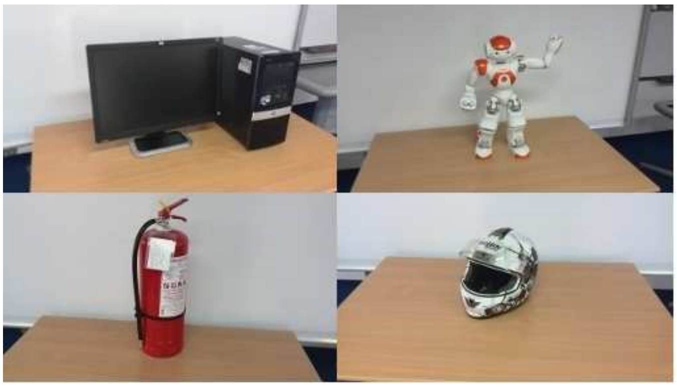


Fig. 1: *Different items used for reconstruction.*

extract information about its structure and label different parts of it. Another fully automated approach is presented in [16] where SIFT features are extracted from an unordered collection of images and used to find matching images of the same scene. Connected components of image matches are calculated and then bundle adjustment is done to solve the camera and structure parameters. Once this is done a 3D model can be generated. This kind of structure from motion reconstruction was applied in [17] and [18]. Their work produced a system for interactively browsing and exploring large unstructured collections of photographs of a scene. This system was made using Bundler, the software which we include in our work in this paper.

III. Experiment Overview

A. Data Collection

We tried to reconstruct the structure of four different items: a monitor and a chassis of a personal computer, a Nao humanoid robot, a fire extinguisher and a motorcycle helmet. These items were placed on a desk in front of a white wall in a brightly lighted room (Figure 1.). They were recorded with the person recording walking along arcs of approximately 180 degrees in front of the items. We did more than one pass, each at a different distance of the items that was in the range between 0.5 and 2 meters. Both the RGB and depth streams of the Kinect were recorded at 640 x 480 pixels resolution at a framerate of around 30 fps. Besides using the Kinect sensor, we captured video recordings using the camera of a Huawei Ascend P6 smartphone [19]. The video of the Ascend P6 was done in a resolution of 1280 x 720 pixels at a framerate of around 30 fps. The quality of the camera of Ascend P6 is moderate compared to the best smartphones available when this work was made, but still it provided noticable higher quality than the RGB Kinect stream and gave us a good measure of the available average smartphone camera. From the video sequences of both the Kinect and Ascend P6 we extracted frame images. We extracted one image per 3-5 frames automatically, and then from the those initial image sets manually we selected the actual images on which reconstruction was to be performed. The video editing, and all following reconstruction work, was done on a computer with Intel Core i3-3217U CPU, 8GB of RAM and Nvidia GeForce GT 635M graphic card.

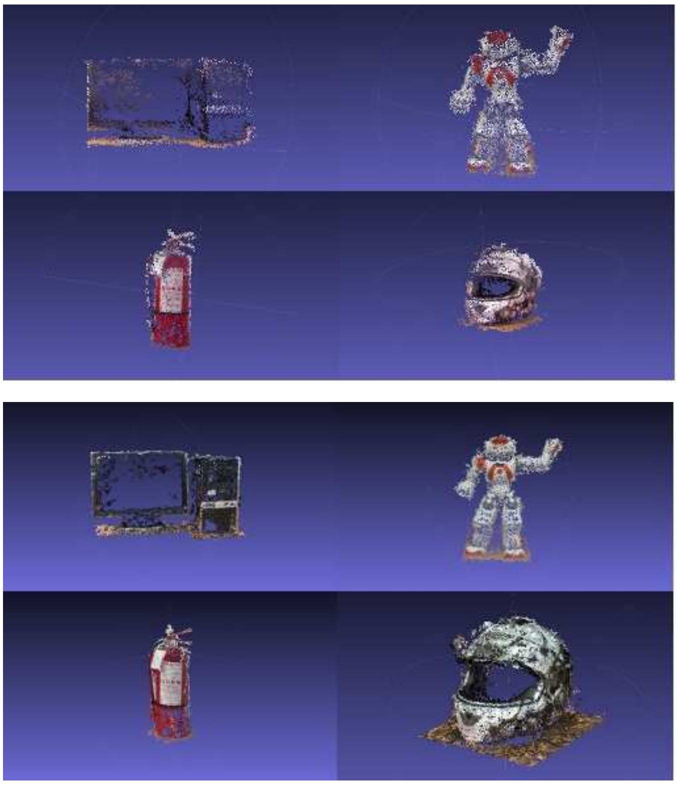


Fig. 2: *Reconstruction results from the Structure from Motion process after applied Patch-based Multi-view Stereo processing. (The top images are generated from Kinect Sensor photos while the bottom ones are generated from photos of the camera of Ascend P6.)*

B. Structure from Motion Reconstruction

For each test item we used a data set of exactly 40 images (with both test cameras). Using a lower number of images gave worse results as expected, and increasing the number of images in most cases gave better 3D reconstructed model, but also incresed the running time of the algorithm. This number was chosen as it gave satisfying results without needed a very long running time. To perform structure from motion reconstruction with Bundler, two camera parameters need to be provided as input - the camera CCD width and focal length in millimeters. The parameters that we obtained and used were 5.954 mm CCD width and 4.884 mm focal length for the Kinect, and 4.800 mm CCD width and 3.979 mm focal length for the Ascend P6. Bundler provides a choice between two image feature extractors - 'siftvlfeat' and 'siftlowe'. In this work while experimenting with different parameters and data sets it was found that 'siftvlfeat' performs better, so all our reported results are done with this feature extractor. The last thing to set before doing reconstruction is the scaling parameter. A value of 1 was used for scaling which means that the pictures were used in their original resolution. After the reconstruction was done the camera parameters were used to get a denser

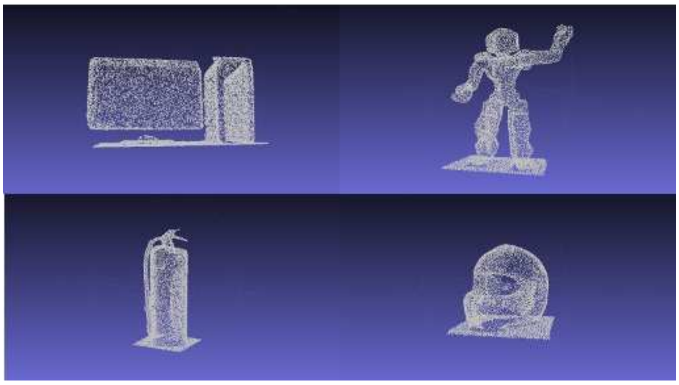


Fig. 3: *Reconstruction results from the Kinect Fusion software.*

TABLE I: Number of points in the different point clouds

Point cloud source	Desktop computer	Nao robot	Fire extin-guisher	Motorcycle helmet
SFM with Kinect Sensor	4.828	4.665	3.287	3.932
SFM with Ascend P6	23.951	10.439	11.880	32.382
Kinect Fusion	17.206	9.687	6.505	7.718

point cloud using PMVS2. Because we used only 40 images, which is a relatively low number, dense reconstruction was done using only PMVS2 without CMVS. The resulting point clouds included noise points which were easily removed using MeshLab [20]. It is a free and open-source software made for processing in the 3D scanning pipeline. The final point clouds from images of both used devices are shown in Figure 2., and the number of points of which the clouds consist are listed in Table 1.

C. Reconstruction with Kinect Fusion

To obtain the ground truth data we used Kinect Fusion [21]. This is a real-time mapping software that can be used in indoor enviroments in various lighting conditions. When using Kinect Fusion to create the 3D mesh the default parameters were used except for the depth threshold which was lowered to 3 meters. The point clouds that we got were very dense, all of the four scenes were on the scale of about one million points. After selecting out the items from the whole scene in MeshLab, we used Poisson disc-sampling to lower the number of points in the cloud. In doing this we did not lose a lot on the detail level of the model, while lowering the complexity of the following tasks to be performed. The results were very accurate and are shown in Figure 3. The number of points in each of the clouds is listed in Table 1.

IV. Results

The first result that we got from our experiment is that doubling the resolution of the photographs on which structure from motion was performed and increasing their quality resulted in greatly increased number of points in the generated point clouds. The denser clouds had almost 5 times more points for the desktop computer reconstruction, double number of points for the Nao robot model, 3.5 times more points for the fire extinguisher and over 8 times more points for the motorcycle helmet. To give further conclusion of our experiment we needed to compare the structure of the point clouds derived from the SFM process with the ground truth data from the Kinect sensor. For this we used the 3D point cloud and mesh processing software CloudCompare [22]. The first step was to bring the two points clouds that we were comparing to the same size scale. Two very distinct points found in both of the clouds that are as much further away from each other were selected and the distance between them was measured. Using this distance we could find the ratio for scaling the point clouds. The cloud that is being measured was scaled up or down to the same dimension of the ground truth cloud. The next step was to do registration of the clouds so that we can minimize the distance between them. To achieve this the Iterative Closest Points (ICP) algorithm was used [23]. After this was done the tool for computing between-cloud distance in CloudCompare was used. The results that were obtained are shown in Table 2. In Figure 4. the reconstructed models from the photographs of the Ascend P6 are shown with their points coloured according to their distance from the ground truth data. These results show that in both cases the reconstructions that were done have the same shape and are good representation of the real life items.

V. Conclusion

In this work we have evaluated the performance of a semi-automated structure from motion process done in an indoor environment. 3D models of different items were reconstructed using images from different camera sources. The results that were obtained prove our initial thesis that even with a low quality camera we can create a 3D point cloud that represents the real life model well. The framework that we used is based on free and open source software so this experiment can be performed easily without a need for some kind of monetary investment, which in time when systems like 3D printing become readily avaialable can be quite compelling. The tests that we did also showed that increasing the quality of the camera considerably increases the quality of the reconstruction. Another way to create a better model is to use a bigger number of images. However, in both cases the execution time of the process will increase. In the future working to automate and improve some of the steps that we did could provide even greater benefit to anyone who would like to create a 3D model of a everyday scene.

Acknowledgements

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje, Skopje, R. Macedonia.

TABLE II: Measured distance metrics between the reconstructed point clouds and the ground truth data

	Kinect Desktop	Kinect Robot	Kinect Fire E.	Kinect Helmet	Ascend P6 Desktop	Ascend P6 Robot	Ascend P6 Fire E.	Ascend P6 Helmet
Min dist.	0	0	0	0	0	0	0	0
Max dist.	0.0618389	0.0802873	0.0823419	0.0487729	0.12322	0.0612998	0.0963241	0.0786461
Mean dist.	0.00978836	0.014143	0.0292939	0.00630127	0.0109317	0.00649334	0.0100716	0.0122751
Sigma	0.00757192	0.0124609	0.0202079	0.00490016	0.0115339	0.0058852	0.00979317	0.0107587
Max relative error	3.77496 + 0.66256/d % (d > 0.0066256)	3.77496 + 0.481724/d % (d > 0.00481724)	3.77496 + 0.525587/d % (d > 0.00525587)	3.77496 + 0.375176/d % (d > 0.00375176)	3.77496 + 0.684554/d % (d > 0.00684554)	3.77496 + 0.510832/d % (d > 0.00510832)	3.77496 + 0.535134/d % (d > 0.00535134)	3.77496 + 0.386784/d % (d > 0.00386784)

References

[1] Autodesk 123D - *http://www.123dapp.com/catch*

[2] 2d3 Sensing - *http://www.2d3.com/*

[3] R. Hartley, A. Zisserman. *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[4] Python Photogrammetry Toolbox and GUI - *http://www.arc-team.homelinux.com/arcteam/ppt.php*

[5] D.G. Lowe, *Distinctive image features from scale-invariant keypoints*, International journal of computer vision, (volume:60 issue:2 p.91–110) 2004.

[6] Bundler - *https://www.cs.cornell.edu/ snavely/bundler/*

[7] M.I.A. Lourakis, A.A. Argyros, *SBA: A Software Package for Generic Sparse Bundle Adjustment*, ACM Trans. Math. Software, (volume:36, issue:1, p.1–30) 2009.

[8] Y. Furukawa, J. Ponce. *Accurate, dense, and robust multi-view stereopsis*, IEEE Trans. on Pattern Analysis and Machine Intelligence, (volume:32, issue:8, p.1362–1376) 2010.

[9] Microsoft Kinect (http://www.microsoft.com/en-us/kinectforwindows/)

[10] K. Khoshelham, *Accuracy analysis of kinect depth data*, ISPRS workshop laser scanning (volume:38, issue:5) 2011.

[11] J. Smisek, M. Jancosek, T. Pajdla. *3D with Kinect*, Consumer Depth Cameras for Computer Vision. Springer London, (p.3–25) 2013

[12] P.E. Debevec, C.J. Taylor, and J. Malik. *Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach.*, Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. ACM, 1996.

[13] F. Schaffalitzky, A. Zisserman. *Multi-view matching for unordered image sets, or How do I organize my holiday snaps?*, Computer VisionECCV 2002. Springer Berlin Heidelberg, (p.414–431) 2002.

[14] S. Teller, M. Antone, Z. Bodnar, M. Bosse, S. Coorg, M. Jethwa, N. Master. *Calibrated, registered images of an extended urban area.*, International journal of computer vision, (volume:53, issue:1, p.93–107) 2003.

[15] A.R. Dick, T.HS Philip, R. Cipolla. *Modelling and interpretation of architecture from several images.*, International Journal of Computer Vision, (volume:60, issue:2 p.111–134) 2004.

[16] M. Brown, D.G. Lowe. *Unsupervised 3D object recognition and reconstruction in unordered datasets*, Fifth International Conference on 3-D Digital Imaging and Modeling, 3DIM 2005, IEEE, (p.56–63) 2005.

[17] N. Snavely, S.M. Seitz, R. Szeliski. *Photo tourism: exploring photo collections in 3D*, ACM transactions on graphics (TOG), (volume:25, issue:3, p.835–846) 2006.

[18] N. Snavely, S.M. Seitz, R. Szeliski. *Modeling the world from internet photo collections*, International Journal of Computer Vision, (volume:80, issue:2, p.189–210) 2008.

[19] Huawei Ascend P6 - *http://consumer.huawei.com/en/mobile-phones/tech-specs/p6-u06-en.htm#anchor*

[20] P. Cignoni, M. Corsini, G. Ranzuglia. *Meshlab: an open-source 3d mesh processing system*, Ercim news, (volume:73, p.45-46) 2008.

[21] R.A. Newcombe, A.J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, A. Fitzgibbon. *KinectFusion: Real-time dense surface mapping and tracking*, In Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on (p.127-136) 2011.

[22] CloudCompare - *http://www.danielgm.net/cc/*

[23] A.W. Fitzgibbon, *Robust registration of 2D and 3D point sets*, Image and Vision Computing, (volume:21, issue:13, p.1145-1153) 2003.

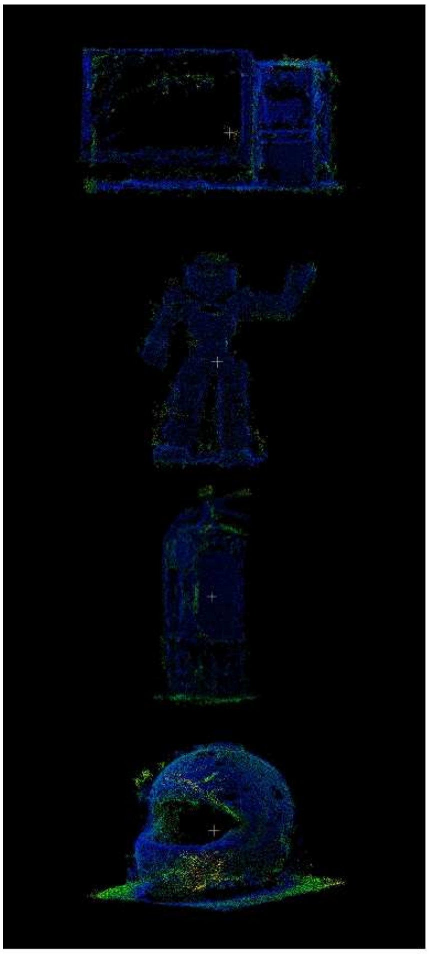


Fig. 4: *The reconstructed models from the Ascend P6 photographs. The color of the points describes their distance from the ground truth data. (blue > green > yellow > red is in the direction of min > max).*

Using machine intelligence techniques in predicting HIV-1 co-receptor usage

Daniel Kareski
Navayo technologies
Skopje, Macedonia
kareski@gmail.com

Nevena Ackovska
Institute of Intelligent Systems, Faculty of Computer Science and Engineering, Skopje, Macedonia
nevena.ackovska@finki.com

***Abstract*—The HIV-1 virus has the ability to infect CD4+ helper T-cells by connecting either to the CCR5 (R5) chemokine co-receptor, CXCR4 (X4) chemokine co-receptor or both. Studies have shown that based on the co-receptor that the virus binds to, we can predict if it has entered its passive or active phase. Machine intelligence tools are used to predict the state of the HIV-1 virus using the V3 loop sequence of the HIV-1 genome thus helping doctors in selecting the right treatment. Different techniques have shown different success rates when it comes to predicting the R5 or X4 affinity. Multiple factors influence the success rate among which the number of test samples used to train the machine learning system and using the right set of key features to distinguish one type from another play a vital role. Also, training the system over test HIV-1 samples which are part of a particular subtype lowers the dispersity of the training and prediction sets so this elevates the precision. In this study a machine learning system used for R5/X4 subtype prediction is presented and the result of its usage.**

Keywords*—*hiv, v3, machine learning, weka

I. Introduction

The HIV-1 is a retrovirus that causes *acquired immunodeficiency syndrome* (AIDS) [1], a condition that weakens the immune system and makes it more vulnerable to all kinds of attacks. This virus has a fast replication cycle which makes it able to modify its structure because of the errors made while self-replicating. Also, its genome is quite compact and efficiently stored due to evolutions way of always using the solution that works the best, so it contains coding regions that overlap. The virus's constantly morphing structure makes it harder for statistical methods to analyze and deduce information which could be used in the development of new anti-HIV medicine. On the other hand, the worldwide AIDS spreading problem is a popular field for developing and testing new bioinformatics prediction methods and techniques. Most of the attempts to stop the virus are directed towards neutralizing its binding mechanism. The HIV-1 virus binds itself to CD4+ helper T-cells using either CCR5 (R5) (shown in fig. 1) chemokine co-receptor, CXCR4 (X4) chemokine co-receptor or both (X4R5). Studing the V3 loop sequence, a subpart of the env gene, shows that there is a high rate of successful prediction regarding the chemokine co-receptor which will be used by the virus. The exact set of features which tell the viruses's co-receptor affinity is still undetermined but great improvement has been made which improves the rate of successful prediction. Another important feature which seems to be related with the C5/X4 prediction is whether the virus is synctytium inducing or not [2]. If the virus induces a cluster of cells which all have a common membrane we call it a synctyium inducing virus. Most of the time, the initial infection of the organism (the virus is in passive state) is followed by non-synctyium inducing R5 virus while in later stages (the virus is in active stage) the virus transforms to synctyium inducing X4 virus. There are many features used in the determination of a C5/X4 subtype which have different effects in different studies and techniques. Most of them are also HIV-1 subtype specific so narrowing down the test set when it comes to subtypes should result in creating a more acqurate prediction. The HIV-1 subject is also popular in Macedonia but there are not many studies conducted with input data from Macedonian patients. This study is a continuation to the 'Comparative analysis of bioinformatics tools used in HIV-1 studies' which was presented on CIIT 2013 and focuses on creating a suitable combination of features which will later be used in creating a machine learning system for co-receptor affinity prediction.

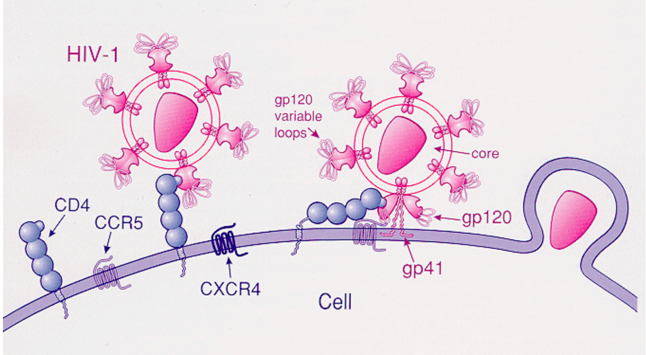


Fig 1. HIV-1 virus using CCR5 co-receptor enters the cell [3]

II. Features Selection

The V3 loop is a part of the envelope glycoprotein gp120 and usually consist of 34-38 base pairs. Initial research of this region suggested that there is a simple rule which shows the co-receptor affinity of the virus. That rule is called 11/25 rule [4] as it just takes into consideration the 11th and 25th position in V3 loop in order to make the co-receptor prediction. The 11/25 rule proposes that a basic amino acid at either the 11th or

25th position of the V3 region is associated with CXCR4 coreceptor usage, whereas acidic or neutral amino acids at these positions are associated with CCR5 coreceptor usage. Introducing new computational methods while performing HIV-1 research showed the inaccuracies produced by the 11/25 rule and both phenotype and genotype experiments revealed that many other features determined whether the virus would bond to R5, X4 or either co-receptor. Besides particular amino acid positions in the V3 loop, in the decision making presented in this paper we also added gap number, net charge and charge rule features. Gap number was included in the training procedure to check if the overall V3 loop length has any effect on the viral phenotype. The net charge is calculated by adding one point for each basic amino acid (Arginine, Lysine, Histidine), subtracting one point for each acidic amino acid (Aspartate and Glutamat), and a neutral value for all other amino acids. The net charge of the V3 loop is reported by multiple studies as a valuable element in the co-receptor prediction [5]. The charge rule is also added as a parameter to evaluate the predictive power of the classificator compared to the conventional 11/25 method and to see if its presence could improve the classificator's performance.

III. Input data preparation

The exported input data which is obtained from the Los Alamos HIV database [6] had to be first reformated in order to be used in the software package. First of all, all the sequences were aligned which means they all were as close to one another as they could by adding gaps or deleting amino acids (this was done using the Mega software package[7]). The input data was formatted as a table with all V3 loop sequence amino acids (40 amino acids after the alignment) put in a different column. At the end of each row we have the three added features (gaps number, net charge and 11/25 rule) and at end, the experimentaly determined subtype of each V3 loop. The dual co-receptor afinity sequences were included in the X4 set which infer that classifiying this data set would show us viruses capable of using the CXCR4 co-receptor opposite to those that are not capable. Out of the selected sequences all the duplicates were removed. The input data after all the preparation changes is shown in table 1 grouped by co-receptor affinity.

TABLE I. Dataset presentation

R5	X4	R5X4
162(62%)	103(38%)	21(8%)

IV. Weka utilization

The bioinformatics software package of choice was WEKA data mining software package [8] because of its reliability and popularity among the data mining community. Modern machine learning (ML) techniques for class prediction can provide advantage over traditional statistics in terms of their abilities to identify and utilize interactions between the chosen features. In addition, the rules some methods generate can often be interpreted with relative ease. The two techniques which are used in this study (C4.5 (J48) tree classifier and support vector classifier) are among the most used tools in bioinformatics, the former for its descriptive power to deduce rules and the latter for its accuracy in classifying and prediction.

A. C4.5 algorithm

The C4.5 algorithm is a decision tree algorithm which can be used for classification. It works by iterating through every unused attribute of the set S and calculates the entropy. It then selects the feature with the least entropy value and splits the data set by the selected attribute (e.g. age < 50, 50 <= age <= 100, age >= 100). The algorithm iterates as long as there are no selected features and in the end each element of the set is either distributed to a class or labelled to the most common class. The output produced by C4.5 is its greatest asset because it represents a human readable format of decision rules. Usually these rules denote that "If particular feature X has value Y than it belongs to class Z". This is especially convenient for machine learning users which do not have extensive informatics experience.

B. Support vector classifier

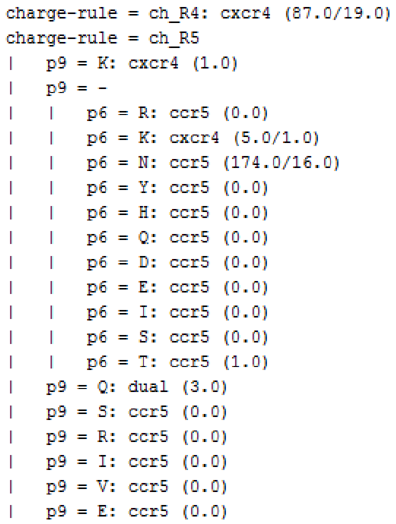
Support vector machines (SVM) are supervised learning models with learning algorithms that analyze data, recognize patterns and are used for classification. The model is trained on a set of training examples where each example is represented by a n-dimensional vector of features and marked as part of one of the two classes. The model is built in such a way that it could assign new examples into one of the two categories. The goal is to map the examples as points in space so that examples which are part of different categories are divided by a clear "as wide as possible" gap. This classifier works best for a two category classification which is a common case in bioinformatics problems. The fact that SVM creates models with high level of confidence makes it a highly used technique.

C. Cross-validation

In order to evaluate the final output of the classifiers a cross-validation method has been used. The k-fold cross-validation (in our example k = 10) is a "hold out" technique for evaluating learning based classifiers. It works by randomly dividing the dataset into 10 subsets with approximately equal size and distribution. Out of the 10 subsets, 9 are used to train the classifier and the last one (the one that is holded out) is used to test the classifier's accuracy. This procedure is repeated for each of the 10 groups. The cross-validation score is the average performance of the classifier. A high score means that the classes are separable to some extent and classifiying a new sample will be successful.

V. Machine learning results

After preparing the input data, it was run in Weka for classifying using both C4.5 and support vector classifiers and 10-fold cross-validation. In both cases Weka's default options were used. After using the C4.5 classifier the sumary showed that a total of 227(83%) examples were classified successfully while 44(16%) were classified wrong. The decision rules tree is shown in fig 2.



```
charge-rule = ch_R4: cxcr4 (87.0/19.0)
charge-rule = ch_R5
|   p9 = K: cxcr4 (1.0)
|   p9 = -
|   |   p6 = R: ccr5 (0.0)
|   |   p6 = K: cxcr4 (5.0/1.0)
|   |   p6 = N: ccr5 (174.0/16.0)
|   |   p6 = Y: ccr5 (0.0)
|   |   p6 = H: ccr5 (0.0)
|   |   p6 = Q: ccr5 (0.0)
|   |   p6 = D: ccr5 (0.0)
|   |   p6 = E: ccr5 (0.0)
|   |   p6 = I: ccr5 (0.0)
|   |   p6 = S: ccr5 (0.0)
|   |   p6 = T: ccr5 (1.0)
|   p9 = Q: dual (3.0)
|   p9 = S: ccr5 (0.0)
|   p9 = R: ccr5 (0.0)
|   p9 = I: ccr5 (0.0)
|   p9 = V: ccr5 (0.0)
|   p9 = E: ccr5 (0.0)
```

Fig 2. C4.5 decision tree

It is fairly easy to interpret the rules displayed in fig. 2. It would translate: “If the charge-rule is ch_R4 then CXCR4, else if charge-rule is CCR5 and p9 is K then CXCR4, ...”. Cases which are not covered by the true are defaulted to CCR5 category. The numbers in parantheses show how many samples reached that leaf versus how many were incorrectly classified. One more important detail showed by WEKA statistics is the receiver operating characteristic (ROC) curve. It shows the ratio of correctly identified positives out of the total positives. In fig. 3 we can see that the last 10 values are almost all false positives.

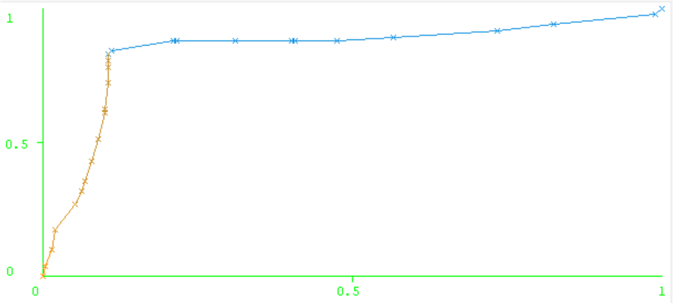


Fig 3. The ROC curve for CXCR4 samples

When using the support vector classifier the summary showed a higher precision of correctly classified samples. Out of the 271 sample, 240(88%) were correctly classified where as 31(11%) were incorrectly classified.

TABLE II. CONFUSION MATRIX

A	B	C	<- classified as
72	2	8	A = CXC4
5	8	8	B = dual
7	1	160	C = CCR5

The confusion matrix shows the number of correctly and incorrectly classified samples for each class specifically.

Considering the fact that we have three classes, the support vector classifier (SVC) works by classifying each combination of two classes and later combines the results. The SVC does not output probability values while performing the classification so its ROC curve is of no use whatsoever. In table III we can see the classification success statistics produced when using this classificator.

TABLE III. DETAILED ACCURACY BY CLASS

TP rate[a]	FP rate[b]	ROC area	Class
0.878	0.63	0.924	CXCR4
0.381	0.012	0.623	Dual
0.952	0.155	0.905	CCR5
0.886	0.116	0.889	Weighted Avg.

[a] True positive rate

[b] False positive rate

We can see that the average TP rates are quite high (in the range of 90%) as opposed to the FP rates which are satisfactory low.

VI. FUTURE WORK

Future work regarding this study would include creating our own classifier. There are lots of studies which offer a better insight in the amount of information each position in V3 loop sequence carries [9]. Also, new features have been extracted which additionally influence the precision of classifying and are used for further deduction of rules and predicting the importance of different V3 loop positions. Some studies even suggest that structural properties and other V loop sequences (for example V2) could have a significant influence on the CCR5 to CXCR4 co-receptor switching of the virus [10].

On the other hand a significant improvement has been made in the field of computing tools. Online machine learning tools are increasing their popularity due to their robustness and practical use. One example of such software is the Ersatz online machine learning software which provides neural-networks-on-the-cloud solution [11]. On-the-cloud solutions provide better computing power but they are limited with customization and the user has to convert the input data to fit the input format of the system. Also, users are unable to modify the software to fit their needs in the most optimal way.

VII. CONCLUSION

Machine learning techniques are a valuable asset when performing bioinformatics data analysis particulary when deducting rules and classifying data on sparse and not extensively numerous data samples. The raw statistical analysis is empowered with different knowledge-gathering techniques that in the end provide simple answers. Both C4.5 and SVM classificators show a good percent of successful classification but customizing these techniques with the newest co-receptor prediction findings could increase the success rate significantly. The WEKA machine learning software is a bioinformatics standard tool for performing such data analysis and provides a great comparison bases when developing new methods for classifying data and data processing techniques in general.

ACKNOWLEDGMENTS

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, R. Macedonia.

[1] Kilmarx P. Acquired immunodeficiency syndrome. In: Heymann DL, editor. Control of communicable diseases manual, 19th Edition. Washington, D.C.: APHA Press; 2008.

[2] Wolfgang Resch, Noah Hoffman, and Ronald Swanstrom, Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence Using Neural Networks, Department of Biochemistry and Biophysics, Department of Microbiology and Immunology, and UNC Center for AIDS Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7295

[3] http://en.wikipedia.org/wiki/CCR5

[4] http://www.prn.org/index.php/management/article/hiv_tropism_1002

[5] Françoise Barre-Sinoussi, Nicole Israël, Christophe Pasquier, Bruno Marchou, Patrice Massip, Pierre Delobel, Marie-Thérèse Nugeyre, Michelle Cazabat and Jacques Izopet, Published Ahead of Print 28 February 2007. J. Clin. Microbiol. 2007, 45(5):1572. DOI: 10.1128/JCM.02090-06.

[6] http://www.hiv.lanl.gov/

[7] MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[9] Wolfgang Resch, Noah Hoffman, and Ronald Swanstrom 2001, Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence Using Neural Networks, Virology 288, 51-62 (2001) doi: 10.1006/viro.2001.1087, available online at http://www.idealibrary.com

[10] Alexandar Thielen, Improve Genotypic Prediction Of HIV-1 Coreceptor usage By incorporating V2 Loop Sequence Variation

[11] http://www.ersatzlabs.com/

Session 8

Intelligent Systems, Robotics, Bioinformatics 2

A Short Survey of Pair-wise Sequence Alignment Algorithms

Done Stojanov, Aleksandra Mileva
Faculty of Computer Science, UGD
Štip, Macedonia
{done.stojanov, aleksandra.mileva}@ugd.edu.mk

Abstract - **We give a short survey of several pair-wise local and global sequence alignment algorithms, together with their comparative analysis. The analysis includes type of the algorithm, its time and space complexity, main characteristics, application for local or global alignment, is the algorithm heuristic or optimal, etc.**

Keywords — Needleman-Wunsch, Smith-Waterson, AVID, MUMmer, BWT-SW, Vmatch

I. INTRODUCTION

One of the oldest and most performed computational task in bioinformatics is sequence alignment. Sequence alignment can be used for multiple purposes, like: annotation of newly sequenced genomes, estimation of evolutionary distances, for producing an approximation and simplification of the history of mutations and evolutionary events, etc. The results from alignment show matching bases, positions where base substitutions has occurred and positions of bases' deletions and insertions. The goal is to find the alignment with optimal matching score. For alignment, a scoring function is used, awarding matches and penalizing mismatches and gaps.

Algorithms for sequence alignment can be optimal or heuristic. The number of possible alignments of two sequences grows exponentially with the lengths of the sequences.

Recently there are many algorithms for short-read alignment against a single reference, which are not subject of this paper.

In this paper we give a comparative analysis of several known local and global pair-wise sequence alignment algorithms. Some good surveys can be found in [7, 16, 23].

II. DEFINITIONS

Let Σ be a finite alphabet. Let A and B be the two sequences that have to be aligned, and let $A = a_1a_2 \dots a_n$ and $B = b_1b_2 \dots b_m$, where $a_i, b_j \in \Sigma$.

For applications in bioinformatics, three most used alphabets are DNA alphabet $\Sigma = \{A, C, G, T\}$, RNA alphabet $\Sigma = \{A, C, G, U\}$, and amino acid alphabet $\Sigma = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$ with 20 letters. So, sequences are DNA, RNA and proteins.

A *global pairwise alignment* S of two sequences A and B results from inserting gaps in A or B or in the both sequences, and after inserting, both sequences are with same length. This means that all letters and gaps in each sequence have to be aligned. Two gaps can not be aligned, because the semantic meaning of a gap is base deletion in the first sequence and base insertion in the other sequence. Global pairwise alignment, can be generalized to *multiple alignment* of a set $\{A_1, A_2, \dots, A_k\}$ of k sequences, which results from inserting gaps in each sequence A_i, so after inserting, all k sequences are with same length. Multiple alignment algorithms are not subject of this paper.

A *local pairwise alignment* S of two sequences A and B of different lengths, identifies local regions of similarity, or matching substrings between them.

Alignment can be seen as simplified representation of the evolutionary history that separates two sequences A and B. Allowed mutations are: simple point mutations, where single character change in the sequence occurs; insertions, where one or more consecutive characters are inserted in the sequence; and deletions, where one or more consecutive characters are removed from the sequence. Other types of mutations, as duplications and rearrangements are also possible, but are not included in the algorithms for sequence alignment.

In order to find the optimal alignment, a score for match, mismatch and gap insertion is defined. One can use PAM approach [11], BLOSUM approach [17] or by using simple scoring function δ which assigns different values, δ_1 and δ_2 for each character match and mismatch in the sequences, and assigns p as gap penalty for each gap. If the gap penalty for the first position of substring of gaps is different from subsequent positions, we speak about *affine gap penalty*, otherwise, if it is same for all positions of gaps, we speak about *linear gap penalty*. If alignment has gaps, it is *gapped alignment*, opposite to *un-gapped alignment*, which is without gaps.

Algorithms for pairwise sequence alignment can be divided in two distinctive classes: optimal and heuristic algorithms. *Optimal algorithms* always find the optimal alignment/s, and these algorithms are deterministic. All optimal algorithms for sequence alignment use the technique of dynamic programming, which is applicable because every partition of the optimal alignment of A and B is the optimal alignment of the corresponding subsequences from A and B. The main characteristic of these algorithms is the quadratic *O(nm)* time complexity. They differ mostly by their space complexity, and best optimal algorithms decreases the space complexity from quadratic to linear.

Heuristic algorithms promise to find reasonable good (suboptimal) alignments, or to find the optimal alignment reasonably often. For the sake of this reduction, they obtain linear time and/or space complexity. These algorithms work in two phases for local alignment. In the first, so called preprocessing phase, matching positions of the highly similar regions $X$and Y are identified as seeds, and in the second phase, the seeds are extended to local alignment and the scoring function is calculated. Usually, not every initial seed is extended to full alignment, instead, many of them are discarded by filtering, which results in the lower runtime. Also, for heuristic algorithms exist a fraction of good alignments for which no seed is found, and they form false negative rate. Bigger false negative rate, means lower sensitivity for the algorithm. Li et al [24] showed that problem of choosing the optimal seed is NP-hard. Brejova et al [6] defined so called vector seed as generalization of many different seeds.

Definition 1 [6] *Vector seed* is an ordered pair $Q = (v, t)$, where $v = (v_1, v_2, \dots, v_l)$ is the seed vector of real numbers and t is the seed threshold value. $X = (x_1, x_2, \dots, x_n)$ hits the seed Q at position $p$if $\sum_{i=1}^{l}(v_i \cdot x_{p+i-1}) \geq t$. The number of nonzero positions of in the vector v is the *support* of the seed.

Similarly as seeded local alignment, exists so called anchor-based global alignment, based on heuristic approach that works in three phases. In the first phase, similar regions called anchors are identified, in the second, some anchors are chosen, and in the third phase, other regions between the anchors are aligned. The quality of the final alignment/s depends on the selection of anchors, and anchors are much more easily selected when the sequences are homologous, i.e. similar sequences with a common evolutionary origin. Two popular seed/anchor-based alignment techniques use hash table and suffix/prefix trie, such as suffix tree [35], enhanced suffix array [1] and FM-index [14]. When hash table is used, in the preprocessing phase, for each seed with length $k$in the sequence A its position in a given array is calculated by hashing function. Afterwards, one can go through the other sequence $B$and find the positions of seeds in linear time. The primary advantage of suffix/prefix tries is that alignment of multiple copies of the same substring is done once because they collapse on unique path in the trie, compared to hash tables, where the alignment is performed for each copy. Suffix trees can be built and searched in linear time and linear space. All three techniques of suffix/prefix tries has the same linear time complexity, but memory requirements differ - suffix tree requires 12-17B per nucleotide, suffix array requires 6.25B and FM-index requires 0.5-2B per nucleotide [23].

Some of the presented algorithms have parallel versions, like MPI-LAGAN [36], which is parallel version of LAGAN [8]. Also, some of the algorithms have their version for multiple sequence alignment, such as LAGAN again.

III. Algoritms and Their Analysis

First algorithms for local and global sequence alignment use dynamic programming for obtaining optimal alignment.

A. *Optimal sequence alignment algorithms*

Needleman-Wunsch pioneering algorithm [27] appeared in 1970 and it is an optimal algorithm for global sequence alignment, based on dynamic programming. It has unfavorable *O(nm)* time and space complexity. Scores are specified by the $(n+1) \times (m+1)$ matrix $S = [S_{i,j}]$ of similarity. The first row is obtained by the formula $S_{0,j} = p \cdot j$, and the first column is obtained by the formula $S_{i,0} = p \cdot i$. The values of the remaining cells are computed by the following formula:

$$S_{i,j} = max \begin{cases} S_{i-1,j} + p \\ S_{i,j-1} + p \\ S_{i-1,j-1} + \delta(a_i, b_j) \end{cases}$$

After completion of the matrix, the highest score is the optimal score, and it can be found in the last cell $S_{n,m}$. The optimal alignment is constructed by tracing back from the last cell to the cell $S_{0,0}$.

Other global sequence alignment algorithms try to reduce the space complexity to be less then quadratic. Hirschberg [18] algorithm is in fact, divide and conquer version of the previous algorithm, and it has linear space complexity. It stores only current and previous row of the Needleman-Wunsch matrix. In each step, the algorithm finds partitioning point (x, y), which divide the two sequences A and B into subsequences $A = A_l A_r$ and $B = B_l B_r$. This point provide obtaining an optimal global sequence alignment of A and B, by concatenating the optimal global alignments of subsequences A_l and B_l, and subsequences A_r and B_r.

Another improvement of Needleman-Wunsch is given by Fickett [15]. It tries to reduce the time complexity by reordering the calculation of matrix values, in order to avoid wasted computation.

Smith-Waterman [32] algorithm is based on Needleman-Wunch approach, without negative values in the cells, generating optimal gapped local alignment. It has the same quadratic time and space complexity. The first row and column of the $(n+1) \times (m+1)$ matrix $S = [S_{i,j}]$ are initialized to zero, and the other values are computed by:

$$S_{i,j} = max \begin{cases} S_{i-1,j} + p \\ S_{i,j-1} + p \\ S_{i-1,j-1} + \delta(a_i, b_j) \\ 0 \end{cases}$$

The optimal local alignment is obtained by tracing back from the cell with the highest score to the first zero cell.

B. *Heuristic pair-wise local sequence alignment algorithms*

Instead of finding one optimal alignment, Waterman and Eggert [34] came up with an idea of identifying k suboptimal local alignments, in *O(knm)* time and quadratic space. Huang and Miller [19] presented linear-space variant of the Waterman-Eggert algorithm.

FASTA [25, 29] is the first seed-based alignment algorithm based on hashing. In the first stage a look up for matching substrings of length k (called hot-spots) is conducted. For indexing hot-spots between query and database sequence, a hash table is employed. Consecutive hot-spots form continuous diagonals in dynamic programming matrix. By choosing 10 best diagonal runs of consecutive hits from the matrix and performing gapped alignment in order to align perfectly matching regions, using variant of the Smith-Waterman algorithm, local gapped alignment is produced. With FASTA, seeds with length less then k could be missed. Since the algorithm employs matrix, the space complexity is $O(nm)$. The average time complexity of the algorithm is $O\left(\frac{nm}{|\Sigma|^k}\right)$.

BLAST (Basic Local Alignment Search Tool) [2, 3] is heuristic approach for searching a database of protein or DNA sequences against target query sequence. A list of similar sequences regarding the query is reported as an output. By breaking the query sequence in overlapping words of size k (the default value for k is 3), for each word at position p in the query a list of words of the same size, scoring at least T with the p-word is generated. During the second phase, the database is scanned for list words' hits. By extending hits in the both directions, until the score of the alignment significantly drops, a local pairwise alignment (ungapped or gapped, depending on whether gaps' insertion is permitted or not during the extension phase) is generated. BLAST has the same time complexity as dynamic programming algorithms $O(nm)$, but since non-significant local alignments are discarded, its running time is about 500 time faster than standard dynamic programming algorithms. Since lookup table of size $|\Sigma|^k$ is stored in the memory, BLAST overall space complexity is $O(|\Sigma|^k + nm)$.

BLAT (BLAST-Like Alignment Tool) [20] is pairwise alignment algorithm that runs about 500 times faster when aligning mRNA/DNA sequences and about 50 times faster when aligning protein sequences, compared to the pre-existing tools. Searching for exact or almost exact hits, BLAT is less sensitive than BLAST. Due to the reduced sensitivity, BLAT is not recommended tool for searching more distantly related sequences, but when it comes to closely related sequences, better time performance results are obtained in comparison to BLAST. During the search stage, three different strategies are used in order to find homologous regions: searching for perfect hits, allowing at least one mismatch between two hits and searching for multiple perfect matches, which are in close proximity to each other.

PatternHunter [26] introduced spaced seeds, building at first an index of A for this model of seeds. Spaced seeds require exact match on k positions, which do not have to be consecutive. It is more sensitive than BLAST and BLAT, with the same time and memory complexity.

BLASTZ [31] is the fastest of all algorithms in BLAST family. Main speedup is obtained by removing all substring repeats in the sequences. BLASTZ uses so called transition seeds of length k, with at most one transition from one to other character, which are extended in both directions without gaps, until score drops below some threshold value t_1. Afterwards, it performs gapped alignments called zones, with score above other threshold value t_2. For the regions between zones, the previous procedure is repeated with smaller value of k and smaller thresholds.

YASS [28] is a variant with tradeoff between FASTA and BLAST. It uses small exact repeats obtained by hashing as seeds and multiple seed criterion that allows an arbitrary number of possibly overlapping seeds. Afterwards, an extension is performed, using new criterion called group criterion, based on the total nucleotide size of the group.

BWT-SW [22] uses FM-index [14] and Burrows-Wheeler Transform (BWT) for emulating suffix trie and obtaining better storage complexity. Authors reported $O(n^{0.628}m)$ time complexity for ungapped alignment. They also stated "how to modify the dynamic programming to allow pruning but without jeopardizing the completeness".

FLAG [33] generates local alignment in linear time and space, when two homologous sequences are aligned. For each overlapping window, the longest matching region is found as a seed, being afterwards extended to local alignment. The algorithm uses a clever way, without have to perform too many unnecessary matching checks for obtaining the "best" local alignment.

C. Heuristic pair-wise global sequence alignment algorithms

MUMmer [12] is the first widely used and efficient anchor-based global alignment algorithm for two genome sequences. It uses suffix tree data structure to find maximal unique matches (MUMs) between A and B, providing $O(n+m)$ time complexity and linear space complexity. MUMs are unique if they occur exactly once in each of the sequences. After finding all MUMs, the algorithm sort them according to their position in the sequences, then picks the longest set of non-conflicting anchors, and aligns the regions between the chosen anchors with Smith-Waterman algorithm. It is open-source now and it has better computational time for homologous sequences, because in that case, Smith-Waterman works less. MUMmer additionally locates all single nucleotide polymorphisms, large inserts, significant repeats, tandem repeats and reversals. Its subsequent two versions are described in [13] and [21]. An additional option of MUMmer 3.0 is the identification of all maximal matches, including non-unique ones, what increases the storage and computational time for creating the output file.

GLASS (GLobal Alignment SyStem) [4] is slower, but more sensitive algorithm than MUMmer. It starts by finding all common K-mers in two sequences. For building alignments, only non-overlapping and non-crossing K-mers with score above some threshold value T are considered. For the regions between the K-mers, the same algorithm is performed recursively for smaller values of K (K takes one of the values 20, 15, 12, 9, 8, 7, 6 and 5). At the end, remaining regions are aligned using standard dynamic programming algorithm.

Another anchor-based global alignment algorithm that uses suffix tree, is AVID [5]. The suffix tree is built for the concatenation of two sequences with a special character N between them. The suffix tree is searched for maximal repeated substrings in linear time. A maximal repeated substring that crosses the boundary between the two sequences represents a maximal match between the two sequences. An anchor set is a collection of non-overlapping, non-crossing matches with length at least half the length of the longest match. The matches are sorted at first, and for selecting good anchors with score above some threshold, a variant of Smith-Waterman algorithm is used, similarly to GLASS. For alignment of regions between the anchors, AVID makes a recursive calls to the same previous process, and previously discarded shorter matches are reconsidered for anchoring in the later rounds. The recursion ends when there are either no remaining bases to be aligned, or there are no significant matches in the remaining sequences. If these remaining sequences are short, they are aligned using the Needleman-Wunsch algorithm, but if they are long, the lack of anchor indicates no significant alignment between them. AVID has good performances only for homologous sequences.

LAGAN (Limited Area Global Alignment of Nucleotides) [8] heuristics use anchors build by chaining an ordered subset of local alignments, which are obtained by seeding strategy with allowed mismatches in the seed. Every local alignment is built from more than one short seeds using CHAOS algorithm [9]. Global alignment is obtained by applying CHAOS recursively in the areas with sparse anchors, so that each consecutive pair of anchors is separated by a distance smaller than a given maximum, which leads to one rough global map. Afterwards, the Needleman-Wunsch algorithm is performed on the limited area around the rough global map [8]. Sensitive anchoring scheme of LAGAN, makes it suitable for close and distantly related sequence pairs.

SPA (Super Pair-wise alignment) [30] is fast global alignment algorithm for homologous sequences, with reduced time and space complexity to $O(m)$, without sacrificing too much accuracy. By measuring the percent of local similarity within shifting window of size k, all positions where nucleotides' deletions occurred can be identified. If by addition of gap/s at concrete position/s, the percent of local similarity within the shifting window is increased over certain threshold, that the insertion of gap/s is permitted.

Vmatch [1] is a variant of MUMmer where suffix tree data structure is replaced with enhanced suffix array. The name enhanced suffix array stands for data structures built of suffix array and additional tables. The time complexity remains the same, but the memory requirements are drastically reduced. Sequence alignment is only one feature of the software Vmatch.

IV. Conclusions

Our analysis is summarized in the Table 1. There are several described algorithms that are not included, due to the fact that their time and storage complexity is difficult to be computed.

References

[1] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch, "Replacing suffix trees with enhanced suffix arrays", Journal of Discrete Algorithms 2, 2004, pp. 53–86.

[2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," Journal of Molecular Biology 215 (3), 1990, pp. 403–410.

[3] S. Altschul, T. L. Madden, A. A. Schoffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", Nucleic Acids Researsch 25, 1997, pp. 3389-3402.

[4] L. Batzoglou, J. Pachter, B. Mesirov, B. Berger, and E. S. Lander, "Human and mouse gene structure: comparative analysis and application to exon prediction," in RECOMB '00: Proc of the 4th Int'l Conference on Computational Molecular Biology, 2000, pp. 46-53.

[5] N. Bray, I. Dubchak, and L. Pachter, "AVID: A global alignment program", Genome research, 13(1), 2003, pp. 97-102.

[6] B. Brejova, D. G. Brown, and T. Vinar, "Vector seeds: an extension to spaced seeds", Journal of Computer and System Science 70(3), 2005, pp. 364-380.

[7] D. G. Brown, "A survey of sequence alignment", Computational Genomics: Current methods. Horizon Press, 2007; N. Stojanovic, ed., pp. 95-120.

[8] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, NISC Comparative Sequence Program, E. D. Green, A. Sidow, and S. Batzoglou, "LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA", Genome Research 13, 2003, pp. 721-731.

[9] M. Brudno and B. Morgenstern, "Fast and sensitive alignment of large genomic sequences", in Proc of IEEE Computer Science Bioinformatics Conference, 2002, pp. 138-147.

[10] M. Burrows, and D. J. Wheeler, "A block-sorting lossless data compression algorithm", Technical Report 124, Digital Equipment Corporation. CA: Palo Alto, 1994.

[11] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins", Atlas of Prot. Seq. and Struct. 5, 1978, pp. 345–352.

[12] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, "Alignment of whole genomes", Nucleic Acids Research 27, 1999, pp. 2369-2376.

[13] A. L. Delcher, A. Phillippy, J. Carlton, and S. L. Salzberg, "Fast algorithms for large-scale genome alignment and comparison", Nucleic Acids Research 30, 2002, pp. 2478-2483.

[14] P. Ferragina, and G. Manzini, "Opportunistic data structures with applications", in Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000), Redondo Beach, CA, USA, 2000, pp. 390–398.

[15] J. W. Fickett, "Fast optimal alignment", Nucleic Acids Res. 12 (1), 1984, pp. 175-179.

[16] W. Haque, A. Aravind, and B. Reddy, "Pairwise sequence alignment algorithms – a survey", In Proceedings of the 2009 conference on Information Science, Technology and Applications (ISTA'09), 2009 pp.96-103.

[17] S. Henikoff, and J. G. Henikoff, "Automated assembly of protein blocks for database searching", Nucl. Acids Res. 19, 1991, pp. 6565–6572.

[18] D. Hirschberg, "A linear space algorithm for computing maximal common subsequences", Communications of the ACM 18(6), 1975, pp. 341–343.

[19] X. Huang, and W. Miller, "A time-efficient, linear-space local similarity algorithm", Advances in Applied Mathematics 12, 1991, pp. 337-357.

[20] W. J. Kent, "BLAT- the BLAST-like alignment tool", Genome Research 12 (4), (2002), pp. 656-664.

[21] S. Kurtz, A. Philippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, , "Versatile and open software for comparing large genomes," Genome Biology 5:R12, 2004.

[22] T. W. Lam, W. K. Sung, and S. L. Tam, "Compressed indexing and local alignment of DNA", Bioinformatics 24, 2008, 791–797.

[23] H. Li, N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing", Brief Bioinform. 11(5), 2010, pp. 473-483.

[24] M. Li, B. Ma, D. Kisman, and J. Tromp, "PatternHunter II: Highly Sensitive and Fast Homology Search", Journal of Bioinformatics and Com-putational Biology 2 (3), 2004, pp. 417–439.

[25] D. J. Lipman, and W. R. Pearson, "Rapid and sensitive protein similarity searches", Science 227 (4693), 1985, pp. 1435–1441.

[26] B. Ma, J. Tromp, M. Li, “PatternHunter: faster and more sensitive homology search”, Bioinformatics 18 (3), (2002), pp. 440-445.

[27] S. Needleman, and C. Wunsch, “A general method applicable to the search for similarities in the amino acid sequences of two proteins,” Journal of Molecular Biology 48 (3), 1970, pp. 443-453.

[28] L. Noe and G. Kucherov, “YASS: enhancing the sensitivity of DNA similarity search”, Nucleic Acids Research, 33, 2005, pp. 540-543.

[29] W. R. Pearson, and D. J. Lipman, “Improved tools for biological sequence comparison”, Proceedings of the National Academy of Sciences of the United States of America 85 (8), 1988, pp. 2444–2448.

[30] S. Y. Shen, J. Yang, A. Yao, and P. I. Hwanq, “Super pairwise alignment (SPA): an efficient approach to global alignment for homologous sequences”, Journal of Computational Biology 9(3), 2003, pp. 477-486.

[31] S. Schwartz, et al., “Human–Mouse Alignments with BLASTZ”, Genome Research 13, 2003, pp. 103-107.

[32] T. Smith, and M. Waterman, “Identification of common molecular subsequences,” Journal of Molecular Biology 147(1), 1981, pp. 195–197.

[33] D. Stojanov, S. Koceski, and A. Mileva, “FLAG: Fast Local Alignment Genereting methodology”, Romanian Biotechnological Letters 18(1), 2013, pp. 7881-7888.

[34] M. Waterman, and M. Eggert, “A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons”, Journal of Molecular Biology 197, 1987, pp. 723 - 728.

[35] P. Weiner, “Linear Pattern Matching Algorithms”, In Proceedings of the 14th IEEE Symposium on Switching and Automata Theory, 1973, pp. 1–11.

[36] R. Zhang, H. Rangwala and G. Karypis, “Whole genome alignments using MPI-LAGAN”, In Proceeding of IEEE International Conference on Bioinformatics and Biomedicine, 2008

.

Table 1. Analysis of local and global sequence alignment algorithms

Algorithm	**Year**	**Characteristics**	**Time complexity**	**Space complexity**	**Local vs global alignment**	**Optimal vs heuristic algorithms**
Needleman-Wunsch	1970	Dynamic programming	$O(nm)$	$O(nm)$	global	optimal
Hirschberg	1975	Dynamic programming with divide and conquer	$O(nm)$	$O(m)$	global	optimal
Smith-Waterman	1981	Dynamic programming	$O(nm)$	$O(nm)$	local	optimal
Waterman-Eggert	1987	Dynamic programming	$O(knm)$ k-number of suboptimal alignments	$O(nm)$	local	heuristic
Huang-Miller	1991	Dynamic programming	$O(nm)$	$O(m)$	local	heuristic
FASTA	1985	Consecutive seeds of length k and hash table	$O\left(\frac{nm}{\lvert\Sigma\rvert^k}\right)$	$O(nm)$	local	heuristic
BLAST	1990	Consecutive seeds of length k and hash table	$O(nm)$	$O(\lvert\Sigma\rvert^k + nm)$	local	heuristic
BLAT	2002	Seeds with allowed 0, 1, or 2 mismatches	$O(nm)$	$O(\lvert\Sigma\rvert^k + nm)$	local	heuristic
PatternHunter	2002	Spaced seeds and hash table	$O(nm)$	$O(\lvert\Sigma\rvert^k + nm)$	local	heuristic
BLASTZ	2003	Removed repeats, transition seeds	$O(nm)$	$O(\lvert\Sigma\rvert^k + nm)$	local	heuristic
BWT-SW	2008	FM-index	$O(n^{0.628}m)$ ungapped alignment		local	heuristic*
FLAG	2013	Extending of the longest consecutive seeds	$O(m)$* homologous sequences	$O(m)$* homologous sequences	local	heuristic
MUMmer	1999	Anchors and suffix tree	$O(n+m)$	$O(n+m)$	global	heuristic
AVID	2003	Anchors and suffix tree	$O(n+m)$	$O(n+m)$	global	heuristic
SPA	2003	Measure of local similarity	$O(m)$	$O(m)$	global	heuristic
Vmatch	2004	Anchors and enhanced suffix array	$O(n+m)$	$O(n+m)$	global	heuristic

Case studies of forest fire detection systems

Igorce Karafilovski
Faculty of Electrical Engineering
and Information Technologies,
"Ss. Cyril and Methodius" University,
1000 Skopje, Macedonia

Vladimir Zdraveski, Dimitar Trajanov
Faculty of Computer Science
and Engineering,
"Ss. Cyril and Methodius" University,
1000 Skopje, Macedonia

Abstract—**The global warming and the frequent forest fires are the greatest evils, that is happening to the world today. They are key motivation factor in development of systems for an early prevention and detection of forest fires. In this paper, we review the general architecture of a forest fire system, present several existing systems and pilot-projects, describe their specific architectures and emphasize the main mutual similarities and differences. At the and we address the open issues and improvement opportunities, that appear with the mobile technologies and smart phones.**

I. Introduction

One third of Europe is now covered by forests, corresponding to 185 million hectares (ha) [1]. The total forest area has increased over the recent decades. The forest fires are one of the greatest evils that are happening in the world today. They are usually accompanied by loss of human lives, loss of homes, loss on forest, biodiversity change and climate change. Most of the damage occur during the summer. Every year forest fires in Europe burn on average about 500 000 ha (twice the area of Luxembourg). Nearly 95 % of the total burned area lies in the Mediterranean region [2]. Therefore, the early prevention and timely dealing with forest fires becomes essential.

The fires are mainly caused by humans, then lightning and other reasons, that contribute to easy start of fires, as the global warming. The prevention of fires is very complex, it requires a lot of work, that includes the long-term process of changing the consciousness of the population and a short-term protection of the forests with a regular maintenance. The short-term aspect is for example, minimizing the side effects including a fire control, increasing the resistance of the assets of a fire, in the case of forests, relocation of resources away from the path of the fire as well as reducing the possibility of a fire occurrence. Although the number of fires in the last decade has increased in Europe, the burned area has not expanded significantly due to the improved firefighting methods.

Several methods and technics are known and used in the world, such as satellites [3], spy planes, surveillance pillars, heat sensors and the combination of these methods. The Satellites are very expensive, only limited areas are monitored and cannot detect small fires. The clouds are also a problem, but in a lack of other solutions for the large inaccessible areas, they are the only way to detect forest fires. The spy planes are used in USA, Canada, Russia, Finland, as well as in other countries with large forest areas, as early warning systems. They are embedded in planes that patrol with regular cameras and use IR-cameras for night vision. The surveillance pillars were used a long time ago as the only possible method that oversees the forest fires.

Lately, these pillars have been complemented with modern and sophisticated cameras or an appropriate types of sensors, that collect information and transmit data to the control centers. Due to their geographical infrastructure, some countries have to use combined methods because they are not able to oversee the whole territory by using only one method.

In the world, there are several systems and pilot-projects, which have been created for an early prevention and detection of forest fires. There are many scientific studies that have covered this issue from different aspects [4] [5]. The starting point in the almost all researches is the great number of forest fires in recent years and the need to act preventive and to deal faster, easier and simpler without major consequences for the population and the forests.

II. Fire System General Architecture

Motivation for analyzing the forest fires early detection and prevention systems are the positive aspects of one such system present in a country. The existence of such a system requires certain conditions to be fulfilled. First, there have to be a good coverage of automated measuring stations, suitable for obtaining meteorological data on the territory that is aimed at protection. Then an access to the data from the satellites is required to calculate the designated parameters. Additionally a detection system in the area is needed, consisted of a combined set of cameras [6], sensors and extended with a crowd-sourcing modules, such as citizens' smart phones. All of these real time data should be merged with the static data (vegetation map, demographic maps, orthophoto maps, etc.) into one integrated system. The system should give all necessary items required for early warning and prevention of forest fires. Fig. 1 shows a general architecture of a Forest Fire System, with a server-side computational unit [7] [8], data providers and users.

III. Example systems

A. Canadian Wildland Fire Information System (CWFIS)

One of the most advanced systems in the world is the Canadian Wildland Fire Information System[1], shown in Fig. 2, that is based on a GIS system. This system creates products on a daily basis, which are maps of the daily fire danger, the annual fire behavior and the fire points (popularly called

[1] Canadian Wildland Fire Information System, http://cwfis.cfs.nrcan.gc.ca

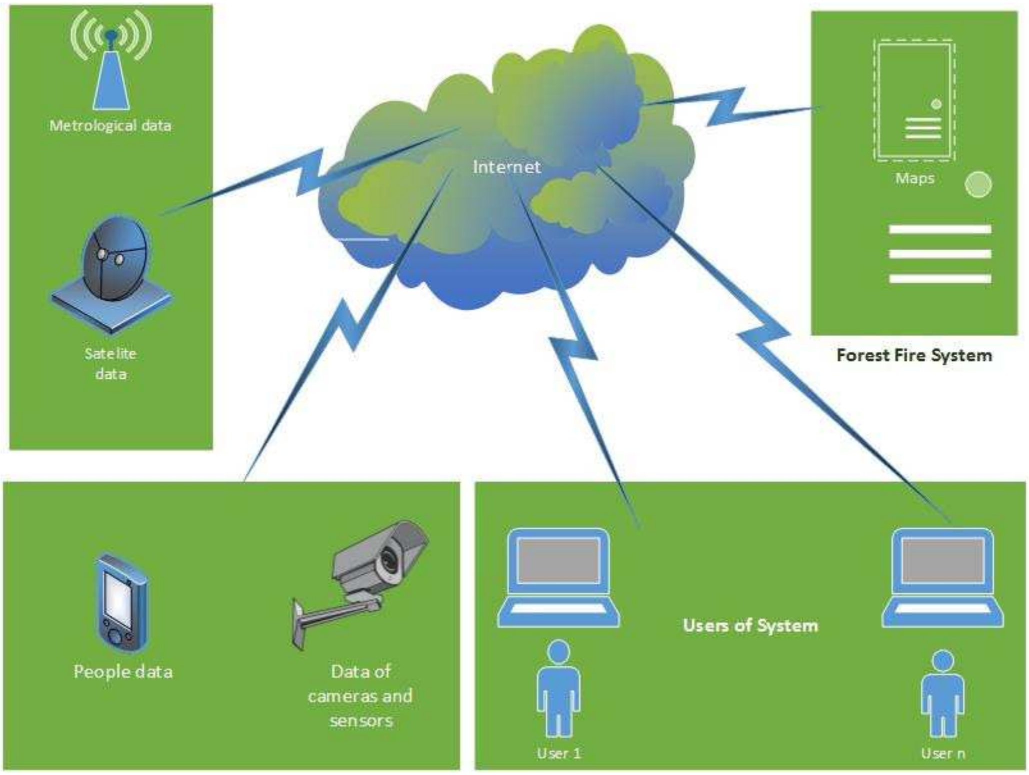


Fig. 1. Forest Fire System General Architecture.

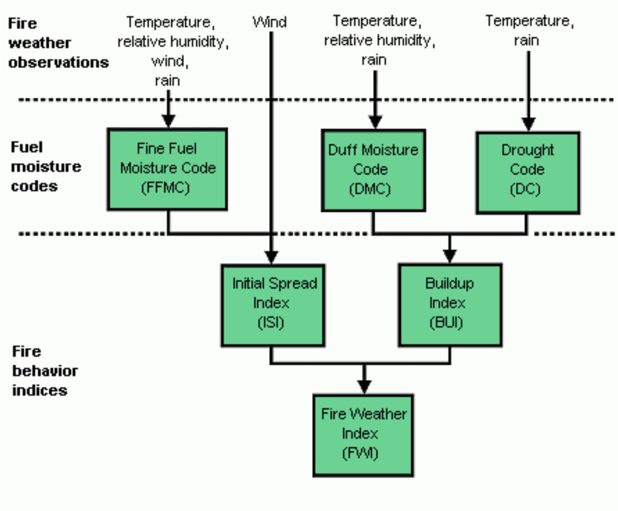

Fire weather observations
Temperature, relative humidity, wind, rain
Wind
Temperature, relative humidity, rain
Temperature, rain
Fuel moisture codes
Fine Fuel Moisture Code (FFMC)
Duff Moisture Code (DMC)
Drought Code (DC)
Initial Spread Index (ISI)
Buildup Index (BUI)
Fire behavior indices
Fire Weather Index (FWI)

Fig. 2. Diagram of Canadian Forest Fire Weather Index (FWI).

“Hot spots”) during the fire season, generally from May to September. These maps are used by the agencies, that deal with fires, for their research and it is open to the public. Forecasts are made six days earlier. This is a computer-based system, that follows the daily danger and conditions for the occurrence of a fire in Canada.

All weather data from Canada is gathered on a daily basis and it is used so that the maps of danger and fire behavior are created. Satellites are additionally used to detect fires. The main benefits of this system are maps, national situational reports for the entire territory and historical analysis. This system consists of several modules, such as:

1) Fire weather data acquisition: The system uses weather data and weather forecasts from the Atmospheric Environment Service (AES) and the Canadian Hydro meteorological Service. The data is gathered automatically and manually from the meteorological stations and the ANIK-D satellite from the North part of Canada. This project integrates 250 meteorological stations on the entire territory of Canada.

2) Data storage and analysis: CWFIS operates in the Workstation County on top of a UNIX platform. The descriptive data model is implemented in the Oracle relational database. The spatial data is stored in ARC/info, in the raster and vector format.

3) Fire Weather Modeling: This system is used by the Canadian Forest Fire Weather Index (FWI) System (Van Wagner 1987), as a basis for modeling and displaying of the possibility of fire occurrence in the forests. The calculation of the components is based on consecutive daily monitoring of the temperature, the relative humidity, the wind speed, and the 24-hour rainfall.

4) Fine Fuel Moisture Code (in this case, the wood, Duff Moisture Code, Duff Moisture Code, Drought Code, Initial Spread Index and Buildup Index): It is the ratings of the total amount of the available combustion fuel, which are numerical values and are calculated by mathematical means.

5) Fire Weather Index: It is the numerical value of the potential fire intensity, that is the combination of the initial spread index and the ratings of the fuel.

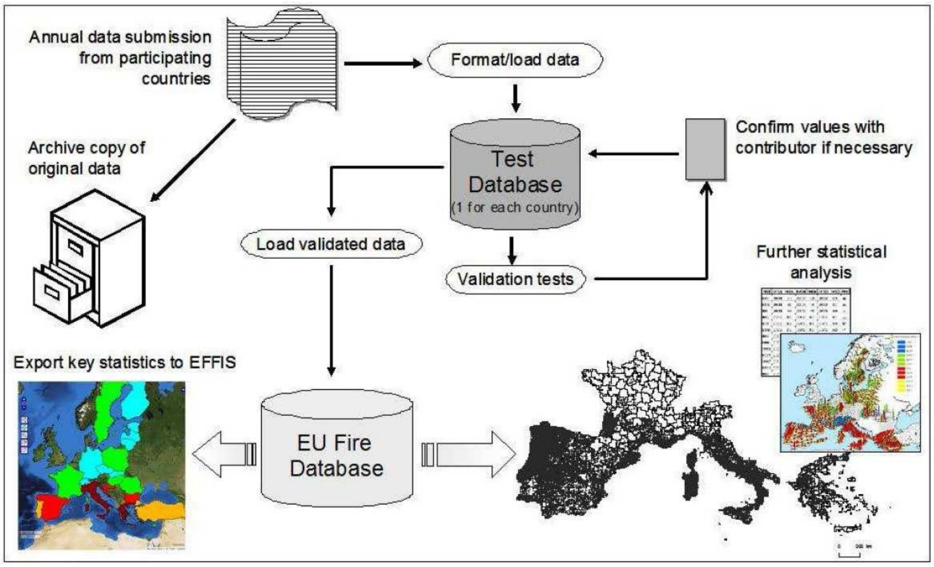

Annual data submission from participating countries
Format/load data
Archive copy of original data
Test Database
(1 for each country)
Confirm values with contributor if necessary
Load validated data
Validation tests
Further statistical analysis
Export key statistics to EFFIS
EU Fire Database

Fig. 3. EFFIS system.

B. European Forest Fire Information System (EFFIS)

EFFIS[2] has been implemented by the Joint Research Centre (JRC) and the General Environmental Directory (ENV) as the main information provider for forests fires in Europe in 1998. The system does evaluation during the two phases, before and after the fire. It takes into consideration the prevention, preparedness, dealing with the fires and the consequences of the fire [1] [9]. Maps of daily meteorological fire danger and a six-day forecast are produced. The satellite images are updated from the last seven days and the latest hotspots maps and possible fires are updated daily.

The whole system, shown in Fig. 3, is divided into two parts: The Current Situation with Fires and Fire Updates.

1) The Current Fire Situation: In this part, EFFIS has presented several products on the map, such as the annual fire forecast, the Hot spot map or the possible potential fires and burned territories. The fire hazard forecast is done with the same methodology that is applied in the Canadian system using the same parameters and calculations. The territory constants are tested and adjusted for Europe because there are parameter differences between Europe and Canada. The data is obtained from two satellites, MODIS [10] and SEVIRI. The information for the burned areas is obtained from all European states including Macedonia, that have an obligation to submit annual reports for the fires.

Technically, the system has been designed as a modular geographic informational system. It consists of web-based modules, a data processing part and a spatial database that collects and shows information for the forest fires on the pan-European scale. The main purpose of EFFIS is to forecast the daily danger of fire and to obtain data for the burned areas, by the use of software tools, meteorological and optical satellite data gathered on a daily basis.

EFFIS works as two inter-dependent systems on on two 64-bit Red Hat Linux servers. These processing (back-end) modules download and process spatial data and generate reports on forest fires. The components ('front-end') consist of web-based mapping tools, through which the EFFIS layers are published and allow the users to search and analyze the information in a web browser.

2) Processing: The spatial and related attribute data are stored in the ORACLE spatial and relational data management system. The MODIS satellite images are saved as ordinary files. Several Payton and Bash Shell scripts, based on GDAL/ORG geospatial libraries, have been developed for processing and management of the raster and vector spatial data, updated on a daily basis or longer. Linux Bash scripts have been developed to download at moderate resolution imaging from the MODIS, TERRA & AQUA satellite data as the images of the German AeroSpace Centre (DLR). The satellite data is merged and a mosaic of 250 meter resolution is created. This data is also used as a mapping base of the rapid assessment of damages, which are used by the experts in forest fires during the fire season.

3) Web-based tools: The EFFIS web-site has been developed in Joomla Content Management System. The web-mapping is the main tool of EFFIS and the most important part is the map search. The browser provides a direct access to FWI as the Web Map Service (WMS), the locations of the hotspots and the burned areas as well as the daily MODIS mosaic. The EFFIS web portal provides access to the news about the fires gathered on the principle of GeoRSS, the data that has been detected daily by the subscribers of the news "plethora of news feed on the web".

GeoServer and UNM Mapserver are both used for management and presentation of the fire danger forecast and other layers associated with the fires in a wide range of formats, including INSPIRE and Open Geospatial Consortium (OGC) standards, such as Web Map Service (WMS), that generates maps in an image format online, Web Feature Service (WFS), which generates vector data using Geographic Mark-Up Lan-

[2] European Forest Fire Information System, http://effis.jrc.ec.europa.eu/

guage (GML) and Web Coverage Service (WCS) that offers raster or grid data.

The future research plans of EFFIS are focused towards the integration of Volunteered Geographic Information VGI and Web 2.0 services. The aim is to include the new resources from the spatial geo-referenced information in the shape of images from the services, such as Flicker and Panoramio, tweets from Twitter and videos from YouTube. These services can provide information and alerts during the fire season because some of these services are relatively new and potential sources of information, in terms of crowd sourcing or crowd sensing. Other researches are aimed towards including meteorological data from sensors using the technology of OCG, Sensor Observation Service (SOS).

C. iForestFire - Croatian intelligent forest fire monitoring system

iForestFire[3], shown in Fig. 4, is an integral and intelligent system for remote monitoring and protection of fire on an open space. An automatic early fire detection is implemented by analysis of camera images in the visible part of the spectrum during the day and of the infrared part of the spectrum during the night.

This kind of a system with 29 cameras has been installed in Istra (Croatia) covering the whole area. As an observer, it has many advantages over the standard observers on the ground. The only disadvantage is that an operator has to watch the screen all the time in order to spot a fire. Therefore, there is a need to automate the fire prediction from the images. Then, the system will automatically recognize the fire and will give a report, i.e. an alarm. The operator will just check and will make a final decision on whether the alarm is true or false. This system uses data from a satellite to make a hotspot map of potential fires just like the Canadian system. iforestFire is a typical web-based information system and the user interface is a web browser. The system consists of terrain unit and a central unit to accept, display, process and store the data from the regional units. The central server unit accepts the data from at most 5 cameras, then processes and displays the results.

It is an integral system and uses three types of data.

1) Video information: The digital video signal is used in the manual and automatic mode of the system. In the automatic mode, the video signal is the source of images for an automatic fire detection. Whereas in the manual mode, the signal is used for a remote video observation and a remote video inspection.

2) Meteorological data: Meteorological data is used in the post-process systems to eliminate false alarms [11]. It also, can be used to calculate the index of forest fire danger during the prevention or in the control phase to follow the spreading of the fire during the fire phases.

3) The Geographic Information System: It contains information that is not only geographical but also other data relevant for the forest fires, such as the fire history, vegetation maps, etc. This part has two working modes, a manual and an automatic mode. In the manual mode, there are control interfaces over the cameras using joysticks, a keyboard, a mouse, virtual controls, geo-referenced maps and panoramic images. In the automatic mode, the system detects smoke in height of 10 meters at a distance of 10 km and automatically activates the alarm.

D. Monitoring forest fires system in the surrounding of Brandenburg, Germany

The Forest Research Institute in the surrounding of Brandenburg (Forest Research Institute, Baden-Wrttemberg) developed a modern system for an early warning and detection of forest fires by the use of Optical Sensor System (OSS). It is also called an automatic monitoring system for the forest fires.

The idea of this type of monitoring came from the German Air Center, where the original software was developed and some parts of the sensors, during the preparation for a space mission. This system was brought into production by IQ Wireless Ltd. and was further tested and developed in cooperation with the German State Brandenburg in 1999. Today, there are 109 cameras for monitoring of the forest areas in Brandenburg. The equipment has been installed on the former control towers, pillars of the mobile transmitters or tall buildings. It can catch the smoke clouds in radius of 15 km with minimum resolution of 10x10 m. The optic sensors can rotate around their own axis and it provides a continuous panorama of 360 °. Three scans (images) are created on every 10 °.

The images are compared and analyzed and the system can show the slightest changes in the atmosphere due to the very fine grey scale that uses over 16,000 nuances. It is a starting point for the origin of the cloud on the electronic map. In this manner, the forest fire can be detected in its early stage (when it starts to smolder). Each system monitors a forest area of about 70,000 ha by the rotation of 4-8 minutes (the setting can be specified if necessary). Currently, there are 11 fire centers where the data is monitored and analyzed. It is planned to be improved and to be merged into 6 modern state fire centers.

About 60% of all the fires have been detected by this system of sensors, whereas the others happen during the late night hours, when the optical sensor system is powerless (blind). Therefore, the system should be extended with a night vision option for detection and analysis.

E. Integrated system for forest fires detection using wireless sensor networks, Waspmote

The company DIMAP-FactorLink, under the name “SISVIA Vigilancia y Seguimiento Ambiental” within the common commercial projects in the protection of the environment, have developed and integrated a system for a forest fires detection that uses the products of Libelium. An area of 210 ha in the region of Northern Spain was covered. The aim was to provide information to different organizations, to monitor the environmental infrastructure and to provide alarms for an early warning of different dangers, such as the forest fires. The system consists of a wireless sensor network [12], a communications network and a reception center [8] [13].

Waspmot devices are deployed in strategic locations as shown in Fig. 5. The four measured parameters are Temperature, Relative humidity, Carbon Monoxide (CO) and Carbon

[3] Intelegent Forest Fire monitoring system, http://ipnas.fesb.hr/index.php

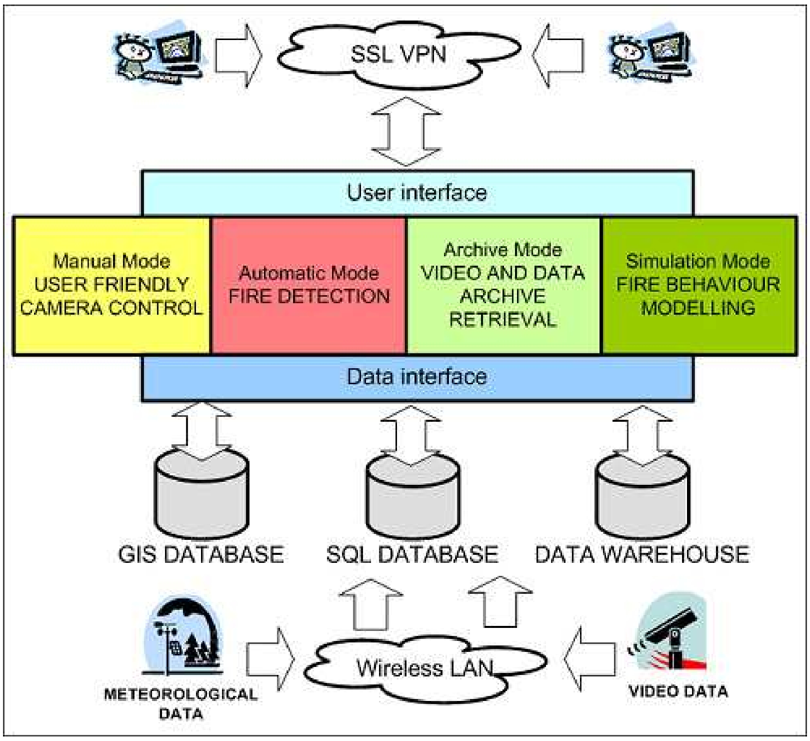


Fig. 4. Croatian iForestFire system.

Dioxide (CO_2). They are measured every 5 minutes. If the value of a measured parameter goes beyond the configured threshold, then the system reacts by sending an alarm to the fire-fighting services. They will immediately know where exactly the fire is, with a reliable accuracy, because within every Waspmote device can be integrated a GPS device that determines the exact position and time of the received information.

One of the main features of the Waspmote is its low power consumption. Waspmote hibernates most of the time in order to save the battery energy. Waspmote awakes on a defined interval (programed by the user), reads from the sensors, establishes a wireless communication and sends data. Every device is powered by batteries and solar panels, that make the system to be completely independent.

Two Meshlium devices are installed to aid the data transmission, to collect information and to send it via WiFi. Meshlium and the Multiprotocol router are capable for an interconnection with WSN (802.15.4 / ZigBee), WiFi (2.4GHz or 5GHz in high or low power), GPRS (quadband, Bluetooth communication with mobile phones or the PDA devices), GPS and Ethernet. The Meshlium device is a parser that divides all the data in small packages or variables that are kept in the server in a MySQL database. The data can be processed after it has been stored in the database. SISVIA has made a control panel to show the information with a graphic interface. The solution has been integrated with GIS, to show the information in 2D or 3D maps.

IV. Conclusions

In this paper were shown five forest fire detection systems and pilot projects, which have been created in different countries and different regions in the world. At the beginning has been analyzed the Canadian Wildland Fire Information System, whose concept is the basis for many other systems, such as the European Forest Fire Information System, shown as a second such system in this paper. In addition we analyzed the Intelligent Forest Fire monitoring system in Croatia, which is a bit different from the previous two, due to its integrated and intelligent video based module for an early detection of forest fires. The system in Brandenburg, Germany, is similar to the Croatian. This is a modern automatic monitoring system for an early warning and detection of forest fires with the use of optical sensors. A different approach for a fire forest detection is the integrated system for a forest fire detection in Spain, that uses wireless sensor networks and Waspmote devices.

The most effective way to minimize the damages caused by the forest fires is the early detection of forest fires and a fast appropriate reaction. In that direction, in the future, more effective forest fire detection systems need to be developed, that will also utilize the new technologies as smart phones. The appearance of smart phones is a good way to use them as a mobile measuring stations and video detection devices. The biggest challenge in the future is to integrate these modern technologies in order to make the forest fire detection and prevention systems more efficient and useful.

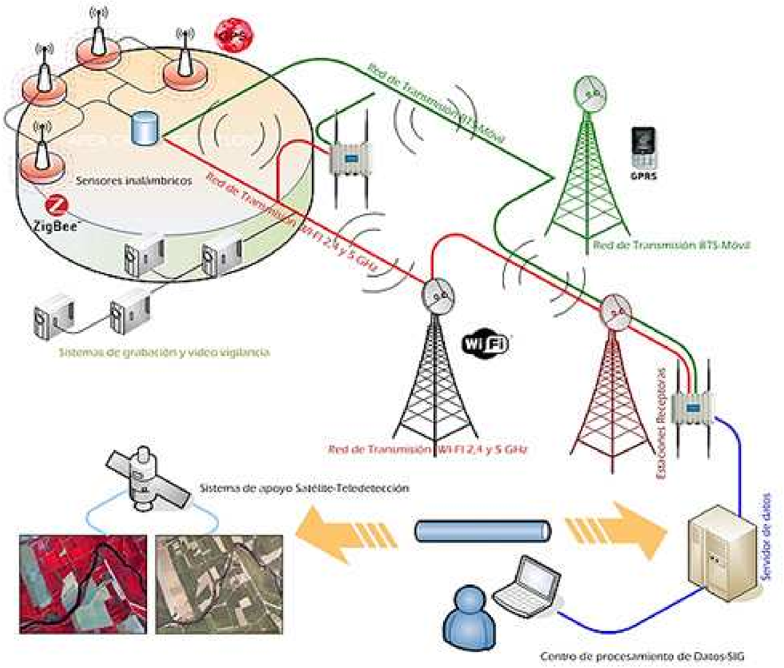

Red de Transmisión BTS Móvil
GPRS
Sensores inalámbricos
Red de Transmisión Wi-Fi 2,4 y 5 GHz
ZigBee
Red de Transmisión BTS Móvil
Sistemas de grabación y video vigilancia
Wi Fi
Estaciones Receptoras
Red de Transmisión Wi-Fi 2,4 y 5 GHz
Sistema de apoyo Satélite-Teledetección
Servidor de datos
Centro de procesamiento de Datos-SIG

Fig. 5. Waspmote system.

Acknowledgment

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the “Ss. Cyril and Methodius” University in Skopje, as a part of the research project “Semantic Sky 2.0: Enterprise Knowledge Management”.

References

[1] J. San-Miguel-Ayanz, E. Schulte, G. Schmuck, A. Camia, P. Strobl, G. Liberta, C. Giovando, R. Boca, F. Sedano, P. Kempeneers, D. McInerney, C. Withmore, S. S. de Oliveira, M. Rodrigues, T. Houston Durrant, P. Corti, F. Oehler, L. Vilar, and G. Amatulli, “Comprehensive monitoring of wildfires in europe: The european forest fire information system (effis),” in *Approaches to Managing Disaster - Assessing Hazards, Emergencies and Disaster Impacts*, ch. 5, InTech, Mar. 2012.

[2] G. Laneve, M. Castronuovo, and E. Cadau, “Continuous monitoring of forest fires in the mediterranean area using msg,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, pp. 2761–2768, Oct 2006.

[3] G. Mazzeo, F. Marchese, C. Filizzola, N. Pergola, and V. Tramutoli, “A multi-temporal robust satellite technique (rst) for forest fire detection,” in *Analysis of Multi-temporal Remote Sensing Images, 2007. MultiTemp 2007. International Workshop on the*, pp. 1–6, July 2007.

[4] K. Fujiwara, K. Kushida, M. Fukuda, and J.-I. Kudoh, “Forest fire detection in far east russian region with noaa avhrr images,” in *Geoscience and Remote Sensing Symposium, 2002. IGARSS ’02. 2002 IEEE International*, vol. 4, pp. 2054–2056 vol.4, 2002.

[5] V. Tsetsos, O. Sekkas, G. Tsoublekas, S. Hadjieythymiades, and E. Zervas, “A forest fire detection system: The meleager approach,” in *Informatics (PCI), 2012 16th Panhellenic Conference on*, pp. 320–326, Oct 2012.

[6] L. Jie and X. Jiang, “Forest fire detection based on video multi-feature fusion,” in *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pp. 19–22, Aug 2009.

[7] A. Chauhan, S. Semwal, and R. Chawhan, “Artificial neural network-based forest fire detection system using wireless sensor network,” in *India Conference (INDICON), 2013 Annual IEEE*, pp. 1–6, Dec 2013.

[8] H.-W. Choi, I.-K. Min, E.-S. Oh, and D.-H. Park, “A study on the algorithm for fire recognition for automatic forest fire detection: The international conference on control, automation and systems 2010 (iccas 2010),” in *Control Automation and Systems (ICCAS), 2010 International Conference on*, pp. 2086–2089, Oct 2010.

[9] D. McInerney, J. San-Miguel-Ayanz, P. Corti, C. Whitmore, C. Giovando, and A. Camia, “Design and Function of the European Forest Fire Information System,” *Photogrammetric Engineering & Remote Sensing*, vol. 79, pp. 965–973, Oct. 2013.

[10] L. Shixing, Z. Yongming, S. Weiguo, and X. Xia, “An enhanced algorithm for forest fire detection based on modis data,” in *Optoelectronics and Image Processing (ICOIP), 2010 International Conference on*, vol. 1, pp. 200–203, Nov 2010.

[11] B. Arrue, A. Ollero, and J. Matinez de Dios, “An intelligent system for false alarm reduction in infrared forest-fire detection,” *Intelligent Systems and their Applications, IEEE*, vol. 15, pp. 64–73, May 2000.

[12] M. Zennaro, A. Bagula, D. Gascon, and A. B. Noveleta, “Long distance wireless sensor networks: Simulation vs reality,” in *Proceedings of the 4th ACM Workshop on Networked Systems for Developing Regions*, NSDR ’10, (New York, NY, USA), pp. 12:1–12:2, ACM, 2010.

[13] J. Zhang, W. Li, N. Han, and J. Kan, “Forest fire detection system based on a zigbee wireless sensor network,” *Frontiers of Forestry in China*, vol. 3, no. 3, pp. 369–374, 2008.

Annotating Protein Structures by Using Multi-label Classifier

Georgina Mirceva, Andrea Kulakov
Department of Intelligent systems,
Faculty of computer science and engineering,
Ss. Cyril and Methodius University in Skopje
Skopje, R. Macedonia
georgina.mirceva@finki.ukim.mk, andrea.kulakov@finki.ukim.mk

Abstract— **The vast amount of knowledge acquired for protein structures hasn't high significance if it is not further used to determine the functions of the proteins in the interactions in which they are involved. Various methods for annotating protein structures found in the literature consider different kind of information about the inspected structures. In this paper we propose a method for determining the protein functions based on the local characteristics of the protein binding sites, as well as the global features of the whole protein structure. First, we extract the characteristics of the amino acid residues that are part of the inspected protein binding site, as well as the features of the protein ray-based descriptor. Then, we build a prediction model by using an existing multi-label classifier. We provide some experimental results of the evaluation of the proposed method.**

Keywords—protein function; protein structure; protein binding site; multi-label classification

I. Introduction

The knowledge regarding the functions of the protein molecules is of high importance, since this knowledge could be used to understand and to regulate various processes in the organisms. There are both experimental and computational methods for annotating protein structures. The experimental methods are expensive and time consuming, therefore there are many protein molecules whose structures are determined but their functions are not yet discovered. Due to this, there is an evident need for development of fast computational methods for annotating protein structures.

There are various computational methods for protein function prediction. First group of methods aims to find homologous proteins structures [1]. Second group of methods tries to find the conserved parts of the proteins sequence and structures in order to determine the protein functions [2]. Third group of methods [3] annotates the protein structures based on the characteristics of the binding sites, which are the regions where the inspected structures comes in interaction with another protein structure. Forth group of methods [4] annotates the protein structures by applying analysis of the protein-protein interaction networks, which contain information about the pairs of protein structures that interacts.

In this paper, we consider the global characteristics of the protein structures, as well as the local characteristics of the binding sites of the proteins molecules in order to determine the functions of the inspected molecule. Then, by using an existing multi-label classifier, we induce prediction model for protein function prediction.

The rest of the paper is organized as follows. In section 2, the proposed method for annotating protein structures is presented. Section 3 contains some experimental results of the evaluation of the method, while section 4 concludes the paper and gives directions for possible further improvements of the proposed method.

II. The Proposed Method

In this paper we introduce a novel method for protein function prediction that is based on the local characteristics of the inspected binding site and the global characteristics of the protein structure. The method uses an existing classifier for multi-label learning for generating the model. The examples that are used in the dataset correspond to the binding sites of the protein chains that are considered for learning and testing the prediction model.

A. Extraction of the Local Characteristics of the Binding Sites

First, we extract the characteristics of the binding sites of the protein structures. Since a given binding site contains several amino acid residues, therefore we calculate the corresponding features for each residue that is part of the inspected binding site. In this research, we consider the Accessible Surface Area (ASA) [5], Relative ASA (RASA), [6], depth index (DPX) [7], protrusion index (CX) [8] and hydrophobicity [9].

The Accessible Surface Area (ASA) [5] holds information about the possibility that a given amino acid residue could be touched by the residues of the interacting protein structure. This characteristic is calculated by using the "rolling ball" algorithm [5]. A rolling ball with a predefined radius is rolled around the surface of the inspected protein structure, and the accessible surface area of each atom is estimated. Each amino acid residue is composed of several atoms that form a particular conformation in the three-dimensional space. By using the rolling ball algorithm, the ASA for each atom is calculated, and then for each residue we calculate its total ASA by aggregating the values of the ASA feature of all atoms that constitute the corresponding amino acid residue.

Amino acid residues contain different number of atoms, thus the ASA feature would be higher for the residues that contain more atoms. Therefore, it is better to use the Relative ASA (RASA) [6], which is a ratio between the estimated value of the ASA feature and the standard ASA value for the inspected amino acid. In this research we use the NACCESS program [6] to calculate the ASA and RASA features.

The third feature that is considered in this research is the depth index (DPX) [7], which is a distance from the inspected atom to the nearest atom that is touched by the rolling probe (has ASA greater than zero). In this way the depth index of each atom is calculated, and then for each amino acid residue we consider the average of the depth indices of all atoms that constitute the corresponding residue.

The protrusion index (CX) is calculated by using the procedure proposed in [8]. For each non-hydrogen atom the number of heavy atoms in its surrounding is calculated. In this paper, we consider the atoms that are on distance no greater than 10 Å, as in [8]. In this way, the inspected surrounding is a sphere with a radius of 10 Å. The protrusion index is calculated as ratio of the non-occupied volume and the occupied volume, where the occupied volume is calculated by multiplying the number of heavy atoms in the inspected sphere and the mean volume of an atom, while the non-occupied volume is calculated as a difference between the volume of the inspected sphere and the occupied volume. In this paper for the mean volume of an atom we use a value of 20.1 Å, as in [8]. In this way the protrusion index of each atom is estimated. Then, we calculate the protrusion index of the amino acid residues by averaging the values of the feature over all atoms.

The hydrophobicity is the fourth feature that we consider in this research, and it is related to the hydrophobic effect of the amino acids. According to this feature we can found out if the inspected amino acid is commonly found in the protein interior or near the protein surface. In this paper we use the hydrophobicity scale proposed in [9].

Since the protein binding sites are formed from several amino acid residues, therefore we calculate the features for each binding site as a sum, average, minimum, maximum and variance of the ASA, RASA, DPX, CX and hydrophobicity features of the residues that constitute the inspected binding site. Also, we consider the number of amino acid residues that form the inspected binding sites as additional feature. In this way we obtain 26 local characteristics of the binding sites. These features are normalized in the interval [0, 1].

B. *Extraction of the Global Characteristics of the Protein Structures*

Besides the local characteristics of the binding sites, we also consider some global characteristics that gives evidence about the conformation of the protein structure in the three-dimensional space. For that purpose, we extract the features of the protein ray-based descriptor that is introduced in our previous work [10]. Next, we present the procedure for extracting the protein ray-based descriptor that is described in [10].

The protein backbone is the skeleton of the protein structure, and its shape holds very important information about the inspected protein structures. The protein backbone is a chain of C_α atoms. In order to provide scale invariance, first the protein structure is scaled, thus the distance between the center of mass and the most distant C_α atom becomes one. The idea of the protein ray based descriptor is to find some representative points along the protein backbone, and then to extract the features as Euclidean distances between these points and the center of mass. However, protein chains have different number of C_α atoms, while we want to obtain descriptors with fixed length in order to be able to make comparison between them. Therefore, we approximate the protein backbone by applying uniform interpolation with some predefined number of interpolation points that are positioned on equal distance along the backbone. In this paper we interpolate the backbone with 64 interpolation points. The interpolation is done is two steps. In the first step, the length of the backbone L is calculated by summing the distances between the consecutive C_α atoms over the skeleton. Then, in the second step, the skeleton is interpolated with N=64 interpolation points that are on equal distances l=L/N along the backbone. Once the skeleton is approximated with fixed number of points, then the features of the descriptor are extracted as Euclidean distances between the interpolation points and the center of mass of the inspected protein chain. In this way, we provide translation and rotation invariance.

C. *Multi-Label Learning Method for Inducing Model*

After extracting the local and global characteristics of the inspected binding sites, next we generate model for annotating the examples with the corresponding functions. The training and testing examples in this case are the binding sites of the protein chains, and they may be associated with multiple functions. Therefore, we have to solve a multi-label learning problem. In [11], the multi-label learning is formulated mathematically, and several methods for solving such problems are described. Also, the most common evaluation measures for estimating the prediction performances of the multi-label learning models are defined.

In the literature there are various methods for multi-label learning. Generally, they can be divided into methods that transforms the multi-label learning problem into one or several multi-class learning problems, and methods that adapt some multi-class classifier for solving multi-label learning problems. In this paper we use the Label Powerset method [11] that transforms the multi-label learning problem into one multi-class learning problem.

Next, we present how the multi-label classification problem is transformed into multi-class classification problem. Let we have a training dataset with four samples $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ that have L different labels (protein functions in this case). In the example described here, we use L=4 different labels denoted as f_1, f_2, f_3 and f_4.

TABLE I. TRANSFORMING THE MULTI-LABEL LEARNING PROBLEM INTO A MULTI-CLASS LEARNING PROBLEM

Example	Original labels	Transformed labels
$\mathbf{x}_1$	$\{f_1,f_3\}$	$m_{\{1,3\}}$
$\mathbf{x}_2$	$\{f_3,f_4\}$	$m_{\{3,4\}}$
$\mathbf{x}_3$	$\{f_2\}$	$m_{\{2\}}$
$\mathbf{x}_4$	$\{f_1,f_2,f_3\}$	$m_{\{1,2,3\}}$

TABLE II. RANKING THE LABELS

Transformed labels m_i	i	$p(m_i\|\mathbf{x})$	member(j,i)			
			j=1	j=2	j=3	j=4
$m_{\{1,3\}}$	{1,3}	0.3	1	0	1	0
$m_{\{3,4\}}$	{3,4}	0.2	0	0	1	1
$m_{\{2\}}$	{2}	0.4	0	1	0	0
$m_{\{1,2,3\}}$	{1,2,3}	0.1	1	1	1	0
$p(f_j\|\mathbf{x})$			**0.4**	**0.5**	**0.6**	**0.2**

The Label Powerset method transforms the multi-label problem into a multi-class problem where the labels in the transformed problem are all unique sets of labels found in the original (before transformation) training data set. Table 1 provide detail about the original and the transformed labels of the samples that are considered in this example.

Let the transformation results into *M* different sets of labels m_i, i=1, 2,…, *M*. Then, by applying some method for multi-class learning, the prediction model is induced. By using this model for a given test sample **x** we can predict the probability $p(m_i|\mathbf{x})$ that the test sample would have the label m_i, i=1, 2,…, *M*. By using the approach presented in [12], we calculate the probabilities $p(f_j|\mathbf{x})$, j=1, 2, …, *L*, that the test sample **x** would have the *j*-th label f_j,

$$p(f_j \mid \mathbf{x}) = \sum_{i=1}^{M} p(m_i \mid \mathbf{x})\text{member}(j,i)$$
$$\text{member}(j,i) = \begin{cases} 1, & j \in i \\ 0, & j \notin i \end{cases}. \tag{1}$$

In this way, the original labels are ranked. The procedure for ranking the labels is described in Table 2. Then, by using some threshold we can define the minimal probability in order to annotate a given test sample with some label. In this paper we use the implementation of the Label Powerset method provided in the MULAN software [11]. For inducing multi-class learning model we consider the C4.5 classifier [13].

III. EVALUATION

A. Data Set Description

Next, we present the procedure for forming the data set used for evaluation of the proposed method. We consider the Gene Ontology (GO) annotations of the representative protein chains that were available on 12 July 2013. The representative protein chains are the chains with less than 100% sequence similarity filtered by the BLASTClust method that is used for clustering the protein sequences based on the distances between them estimated by the BLAST method [14]. From the representative protein chains we filter the chains with less than 30% sequence similarity by using BLASTClust, thus obtaining the chains for testing the prediction performances of the models, while the remaining representative chains are used for training the models. Since in this research we make predictions about the functional annotations of the protein binding sites, therefore we consider only the protein chains whose binding sites are stored in the Biomolecular Interaction Network Database (BIND) database [15]. Then, we consider the chains that have at least one annotation with the functions that are present among the annotations of the training and testing chains. In this way we obtain 2136 training chains and 960 chains for testing. In the data set used for learning and testing the models, the samples are the binding sites of these chains. In this way we obtain 3167 training samples and 1449 test samples. The number of different labels (functions) is *L*=757.

B. Evaluation Measures

In this research we are solving multi-label learning problem, thus we must use evaluation measures that are appropriate for estimating the prediction performances of multi-label classification models. In [11], the most commonly used evaluation measures for multi-label learning are presented.

In this paper we consider several example-based measures and several label-based measures defined in [11]. From the example-based measures we consider Precision, Recall, F_1 and Accuracy, which are calculated as in [11]

$$\text{Precision} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{|A_i \cap P_i|}{|P_i|} \qquad \text{Recall} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{|A_i \cap P_i|}{|A_i|}$$
$$F_1 = \frac{1}{Q}\sum_{i=1}^{Q}\frac{2|A_i \cap P_i|}{|A_i|+|P_i|} \qquad \text{Accuracy} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{|A_i \cap P_i|}{|A_i \cup P_i|}, \tag{2}$$

where with A_i and P_i we denote the sets of actual and predicted labels for the *i*-th test sample, $|S|$ is the size of a given set *S*, while with *Q* we denote the number of samples used for testing the models.

Besides the evaluation measures based on averaging over all examples, also there are measures that average some evaluation measure over all labels (functions). Let BM(*TP*, *TN*, *FP*, *FN*) is some evaluation measure that could be used for estimating the prediction power of some model for solving binary problems (problems with 2 output classes), where with *TP*, *TN*, *FP* and *FN* we denote the number of true positives, true negatives, false positives and false negatives, respectively. In this paper we consider the Precision, Recall, F_1 and Area Under the ROC Curve (AUC-ROC) measures. Then, we calculate the macro and micro versions of this measures as in [11]

TABLE III. EXPERIMENTAL RESULTS BY USING VARIOUS CHARACTERISTICS

Evaluation measure	Local	Global	Local and global
Precision	0.061	0.113	0.113
Recall	0.063	0.118	0.125
F_1	0.055	0.108	0.111
Accuracy	0.039	0.092	0.096
$\text{Precision}_{\text{macro}}$	0.021	0.063	0.066
$\text{Recall}_{\text{macro}}$	0.019	0.106	0.097
$F_{1\ \text{macro}}$	0.016	0.068	0.069
$\text{AUC-ROC}_{\text{macro}}$	0.482	0.565	0.551
$\text{Precision}_{\text{micro}}$	0.057	0.113	0.120
$\text{Recall}_{\text{micro}}$	0.066	0.130	0.131
$F_{1\ \text{micro}}$	0.061	0.121	0.125
$\text{AUC-ROC}_{\text{micro}}$	0.502	0.570	0.564

$$\begin{aligned} \mathrm{B}_{\text{macro}} &= \frac{1}{L}\sum_{i=1}^{L}\mathrm{B}(TP_i, TN_i, FP_i, FN_i) \\ \mathrm{B}_{\text{micro}} &= \mathrm{B}\left(\sum_{i=1}^{L}TP_i, \sum_{i=1}^{L}TN_i, \sum_{i=1}^{L}FP_i, \sum_{i=1}^{L}FN_i\right), \end{aligned} \tag{3}$$

where TP_i, TN_i, FP_i and FN_i are the numbers of true positives, true negatives, false positives and false negatives for the i-th label. L denotes the number of different labels used in the data set.

C. Experimental Results

In this paper we make experiments by using the local characteristics or global characteristics, and by using both local and global characteristics. In Table 3 the experimental results are presented by using various features. From the results we can see that by using only the local characteristics the model obtains significantly lower results than by using only the global characteristics. Generally, it is better to use both local and global characteristics in order to increase the prediction power of the model. The best model is obtained by using all the features, and it has micro precision of 0.120 and micro recall of 0.131. We want to mention that in multi-label learning the evaluation measures are calculated in different manner than in multi-class learning, thus by using numerous labels most of which are very rare, it is expected that the evaluation measures would obtain such values for the evaluation measures.

IV. CONCLUSION

In this paper we introduced a novel method for annotating protein structures based on the local characteristics of the protein binding sites and the global characteristics of the entire protein structure. By using the Label Powerset method, which transforms the multi-label learning problem into multi-class learning problem, we induced models by considering various features. The experimental results showed that the global characteristics hold more relevant information for predicting the protein functions. However, the best option is to combine the local and global characteristics in order to obtain more accurate model.

In the future, we can incorporate some additional characteristics that could provide more relevant information about the functions of the protein structure. Also, some other multi-label learning method could be applied in order to induce more accurate models for protein function prediction.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the “Ss. Cyril and Methodius University in Skopje”, Skopje, R. Macedonia.

REFERENCES

[1] A. E. Todd, C. A. Orengo, and J. M. Thornton, “Evolution of function in protein superfamilies, from a structural perspective,” J. Mol. Biol., vol. 307, no. 4, pp. 1113–1143, 2001.

[2] A. R. Panchenko, F. Kondrashov, and S. Bryant, “Prediction of functional sites by analysis of sequence and structure conservation,” Protein Science, vol. 13, no. 4, pp. 884–892, 2004.

[3] N. Tuncbag, G. Kar, O. Keskin, A. Gursoy, and R. Nussinov, “A survey of available tools and web servers for analysis of protein-protein interactions and interfaces,” Briefings in Bioinformatics, vol. 10, no. 3, pp. 217–232, 2009.

[4] M. Kirac, G. Ozsoyoglul, and J. Yang, “Annotating proteins by mining protein interaction networks,” Bioinformatics, vol. 22, no. 14, pp. e260–e270, 2006.

[5] A. Shrake and J. A. Rupley, “Environment and exposure to solvent of protein atoms,” Lysozyme and insulin, J. Mol. Biol., vol. 79, no. 2, pp. 351–371, 1973.

[6] S.J. Hubbard and J.M. Thornton, NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London, London, UK, 1993.

[7] A. Pintar, O. Carugo, and S. Pongor, “DPX: for the analysis of the protein core,” Bioinformatics, vol. 19, no. 2, pp. 313–314, 2003.

[8] A. Pintar, O. Carugo, and S. Pongor, “CX, an algorithm that identifies protruding atoms in proteins,” Bioinformatics, vol. 18, no. 7, pp. 980–984, 2002.

[9] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” J. Mol. Biol., vol. 157, no. 1, pp. 105–132, 1982.

[10] G. Mirceva, S. Kalajdziski, K. Trivodaliev, and D. Davcev, “Comparative Analysis of three efficient approaches for retrieving protein 3D structures,” 4-th IEEE Cairo International Biomedical Engineering Conference 2008 (CIBEC ’08), Cairo, Egypt, pp. 1-4, 2008.

[11] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining multi-label data,” Data Mining and Knowledge Discovery Handbook, 2nd ed., O. Maimon, L. Rokach, Eds. Springer, 2010, pp. 667–685.

[12] J. Read, “A pruned problem transformation method for multi-label classification,” In Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008), 143–150, 2008.

[13] R. Quinlan, “C4.5: Programs for Machine Learning,” 1st ed., Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.

[14] S. Altschul, W. Gish, W. Miller, E.W. Myers, and D. Lipman, “Basic local alignment search tool,” J. Mol. Biol., vol. 215, no. 3, pp. 403–410, 1990.

[15] G.D. Bader, I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, and C.W. Hogue, “BIND: the Biomolecular Interaction Network Database,” Nucleic Acids Res., vol. 29, no. 1, pp. 242–245, 2001.

Big Data dynamic visualization based on user behaviour

Dimitar Jovanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
dimitar.jovanov.92@gmail.com

Aleksandar Stojmenski
Faculty of Computer Science and Engineering
Skopje, Macedonia
aleksandar.stojmenski@gmail.com

Riste Stojanov
M.Sc., Faculty of Computer Science and Engineering
Skopje, Macedonia
riste.stojanov@finki.ukim.mk

Dimitar Trajanov
PhD, Faculty of Computer Science and Engineering
Skopje, Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*— The amount of interconnected data on the web is rapidly growing, together with the information hidden into it. In order for the important aspects of the data to be revealed, it needs to be analyzed. We, people, tend to visualize the knowledge in order to understand it better. Most of the current tools for visualization are displaying all of the data at once, making it difficult to extract any conclusions from the overwhelmed presentation with many connections. In this paper we propose dynamic and partial graph data visualization, using heuristics and user behavior analysis in order to extract the most important concepts and the relations among them. We are maintaining user context to capture the behavior, and based on it we extract the most relevant concepts that should be displayed. In the prototype system that we have designed as a proof of concept, the amount of displayed data is automatically determined based on the display panel size, in order to avoid over - crowded visualization. The system is providing modular heuristics definition and it is evaluated with several predefined standard algorithms.**

Keywords—big data visualisation; graph representation; user behaviour; javascript

I. INTRODUCTION

It is impossible to visualize all of the data on a small screen. Even if we can, it is hard to extract information from the over-crowded presentation. Only the part of the knowledge that is of user interest should be displayed. Most of the user interests are dynamic and change over time, so the visualization of data should follow them. A visualization program is analogous to a looking glass through which the user inspects an underlying system. In other words, it is the “bringing out of meaning in data” [1]. Users search and navigate through the knowledge of their interest (behaviour). The required data is close to their context and behaviour. The proposed approach consists of three steps: assessing the relevance of nodes, reducing the specification, and presenting the results.

In the following paper, algorithms and methods are presented for these steps along with examples. We use the d3.js library as one of the good data visualization tools for graph visualisation out there. It provides a variety of new ways to visualize data and by working on the web it provides operating system interoperability, because all a user needs to see d3 visualizations is a web browser.

The library used for the display of those data dimensions is d3.js[1]. The d3.js library was chosen because its ability to manipulate all of the DOM elements and exploits most of the browsers built in functionalities while facilitating mouse interaction. Some of its disadvantages are that DOM manipulation can be extremely slow for large numbers of entries. SVG also has performance limitations when dealing with large quantities of elements. In order to provide useful graph visualization, the number of nodes should be related to the available presentation area, which significantly reduces the number of nodes, and most of the visualization libraries don’t have problems with manipulation with this many quantities of data.

II. RELATED WORK

Researchers have previously identified comprehension issues with visualization of big data schemes. Efforts have been made to simplify this process using partial graphs visualisation. Partial graphs are graphs where not all nodes can be reached from the starting condition, or the end condition is not reachable by any node, or both. In practice the end condition is typically unreachable in a partial graph. Partial graph visualization techniques aim at producing graph layouts that are readable, interpretable, and look aesthetically pleasing to the viewer [2]. A graph structure that changes over time is called a dynamic graph. Dynamic graphs are often shown as a sequence of single images put next to each other. Another approach is to connect the diagrams of the subsequent graphs more closely by integrating them into a single diagram and aligning multiple visual representations of the same vertex or

[1] http://d3js.org/

edge over the entire sequence of graphs [3]. Various graph visualization tools were developed by researchers including Blast2GO[4], Neo4J[5] and D3JS[6]. Those tools enable interactive visualization and exploration for all kinds of networks and complex systems, dynamic and hierarchical graphs. Due to an increase in the amount of information being analysed, the number of graph nodes and connection edges increases rapidly. The end result is graph nodes and edges “overlapping”, which complicates the perception of the displayed information for the user. In this case, a possible solution for this non-trivial problem is a usage of algorithms for positioning of graph models [7]. There are many algorithms of such kind, which form different categories, but usage of these is narrowed as it depends on context of information being visualized. A straightforward algorithm for laying out the other nodes, called “radial drawing” in [8], is used in literature. Nodes are arranged on concentric rings around the focus node. Each node lies on the ring corresponding to its shortest network distance from the focus. Immediate neighbours of the focus lie on the smallest inner ring, their neighbours lie on the second smallest ring, and so on. Our implementation draws these rings explicitly to make the network distance apparent. The angular position of a node on its ring is determined by the sector of the ring allocated to it. Each node is allocated a sector within the sector assigned to its parent, with size proportional to the angular width of that node’s sub tree. A method similar to this is described in some detail as “radial placement” in [9], where all the nodes are the same size, and so the angular width of a node’s sub tree is simply the number of leaf nodes among its descendants scenario.

III. User context and behaviour

Visualizing dynamic directed and weighted graphs with an additional hierarchical organization of the graph vertices is a challenging task. Many data dimensions have to be represented at the same time:

- The graph vertices
- The adjacency edges induced by the graph
- The weights of the adjacency edges
- The inclusion edges induced by the hierarchy
- The evolution of the graph over time

Besides these aspects, in this paper, we start with the premise that the user have some information of interest that should be extracted from a large amount of data that can’t be presented on the available displaying area i.e. the goal of the user. Also, during the time, the interest of the user can change, while the goal is steal the same, and the visualization algorithm should also consider this aspect. The visualized data should be the subset of dataset that is “closer” to the goal and the interest of the user. In this paper is presented a prototype system that manages the context of the user and visualizes a subset of the dataset that is “closer” to the user context.

The user context is represented with two main types of nodes. The first ones are the fixed nodes and they describe the constants in the scenario, including the user data, if there are data representing it, as well as the goal in that scenario. In our system, the fixed nodes are preconfigured and are marked as fixed explicitly. The second type of nodes are the temporal nodes, and they represent the recent interest of the user, obtained by monitoring of the user interaction with the system. The interest and the focus of the user can change over the time, and in our system, the more recent actions have a higher priority. In the prototype system, when the user interacts with some node[2], that node is added as temporal in the context.

```
visualize(context, displayNodesNum) {
      nodeContextDist ={};
      init res[displayNodesNum];
      foreach(node in nodes){
        nodeContextDist[node] =
              dist(node, context, res);
      }
      return res;
}
```

The visual nodes should help the user to achieve the goal, and they are presented to the user. The visual nodes are subset of all of the nodes and are selected depending on their distance from the context. The visual nodes are recomputed whenever the context is changed, which is when the user interacts with some node. In this case, the system goes through all of the nodes, computes their distance from the context nodes, and displays the most important ones. The algorithm for this procedure is the following:

The visualize function takes the context object that holds the fixed and the temporal nodes, and the number of nodes that should be displayed. The number of nodes is computed based on the available area for displaying and the size of the node, as shown in (1). The *nodeContextDist* variable holds the distance between each node and the context, computed by the *dist* function. The *dist* function is configurable and can be changed, and the version used by our prototype is shown in (2). After all the distances between the context and all nodes is computed, the closest *displaysNodeNum* nodes to the context are returned. The *res* variable is sorted array that holds the closest *displayNodesNum* nodes. The *dist* function adds new node in the *res* array when it is closer than the last node in *res,* or when there are less than *displayNodesNum* elements in the array.

IV. Data of interest determination

The number of nodes that are displayed is defined in (1), where w_c, h_c, w_n and h_n are the width and the height of the

[2] Node double click action is used to represents the interaction

displaying area and the node elements correspondingly, while α is the weight factor that represents the amount of the displaying area occupied by nodes.

$$displayNodeNum = \alpha * \frac{w_c{*}\mathrm{h}_c}{w_n{*}\mathrm{h}_n} \qquad (1)$$

The weighting factor α is configurable and is initially set to 0.25 to provide a lesser congestion factor between the nodes.

Since the focus of attention are a few nodes from the fixed context, it is natural to place these focus nodes at the centre of the display and layout the other nodes around it. The temporal nodes can be managed using different paging algorithms, such as FIFO, LRU, SJF, but currently we are using FIFO for managing of the temporal nodes because of simplicity. The selection of the visible nodes is made using our most important node first algorithm. The nodes are ranked based on their proximity (number of hops) and relatedness (number of connections) to the fixed and temporal nodes. More specifically, each node has its own score of the metric defined which is as follows:

$$WF * \Big[\sum_{i \in FixedNodes} nc(fn_i, node) + \sum_{j \in TempNodes} nc(tn_j, node) + \sum_{i \in VisibleNodes} nc(kn_i, node)\Big] \qquad (2)$$

The formula displayed above explains the score of each node while determination the visibility of the node. The function *nc* counts the number of connections between two nodes passed as arguments and is used to calculate the distance between the nodes. The arrays *fn*, *tn* and *kn* represent the arrays of fixed nodes, temporal nodes and visible nodes respectively. Each of the elements of these arrays is sent to the number of connections function defined above along with the parameter node (the current node that we are calculating score of). *WF* is a constant that is used as a weighting factor and currently is set to 0.25. After each change of the visible nodes (due to a click or change in the fixed/temporal nodes) every node gets a new score based on the current fixed and temporal context. The first x nodes with highest score are displayed and marked as visible where x is determined by the size of the screen in which the data is being visualised.

V. Evaluation

The system was tested with generated dataset of the university field, containing the relations between the students and the professors. However, the approach is general and applicable to all graph structured data. In our example this would lead to the classes that the current professor teaches, his colleague professors and teaching assistants that teach the same subjects and the students that attend his classes. If we further choose one of the subjects, this leads to a second iteration of our algorithm which produces just the teaching assistants that teach the given subject and are assigned to the particular professor, furthermore all of the students that take this class and attend his lectures are displayed. If for instance we change the chosen subject and pick a different one, the system takes in consideration the previously selected one in calculating the importance factor for the new nodes. This results in students that have attended both classes showing up as more relevant than some of the current students. This in term results in a learning curve that makes the system oriented more towards the user and better at providing fast and accurate data from the enormous dataset.

Another example is the visualization of airline companies, airports and the flights between them. If we fix two of the airports, this will result in a dataset containing all of the flights between them and the airlines that operate with them. Then by choosing two airline companies the flights between the two airports get filtered to only those who belong to the given airlines, if we then deselect one of them the newly filtered data that belongs to both of the companies will have a higher priority, than the ones that belong only to the one that is left selected. Providing more iterations lead to a better understanding of the user and the dataset that he would prefer to see over the unnecessary data that he would get without the system's learning side.

Using the above-defined rules for the number of nodes based on the size of the user's viewport, we could provide the following results:

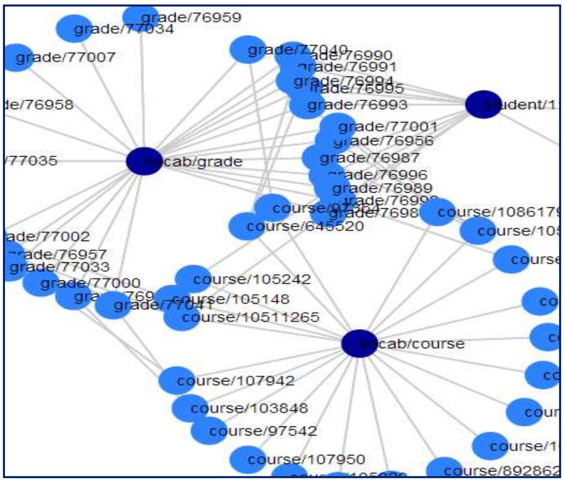


Fig. 1. Nodes per screen configuration.

If the weight factor is set to 0.75 the nodes become more tightly bound, they cause a greater overlap between each other and the whole graph layout loses its context. This can be seen in the following picture:

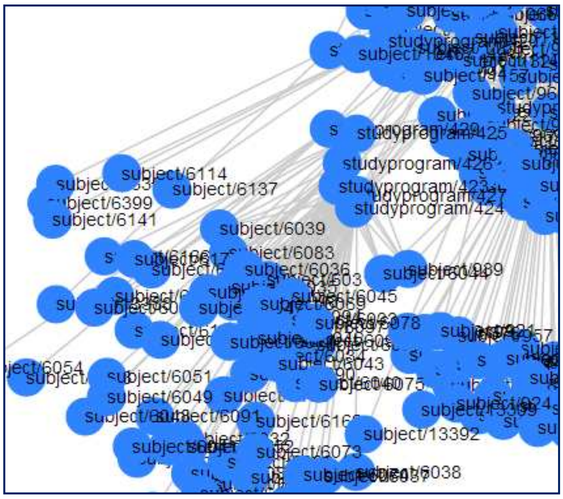


Fig. 2. Nodes with higher weight factor.

For node replacement we tried a variety of combinations including the first in first out algorithm, least recently used algorithm and shortest job first algorithm. Due to it's easy to implement nature, the first in first out algorithm provided fastest results, but the problem was that this particular implementation strayed far from the user aspect of our system. Because following the importance of nodes for the user proved to be very hard. Least recently used provided much better results if the user had used the application for a while, but did very poorly when the user was beginning to interact with the system. Shortest job first or in our case most important node first provided the best results in displaying the nodes that the user wanted to see but remembering the nodes important for the user that had passed was somewhat of a challenge.

Best results, in displaying the designated nodes without speed performance consideration, were achieved with a combination of both shortest job first algorithm and the least recently used algorithm. Shortest job first was used to determine the importance of nodes and least recently used was needed to filter the nodes that weren't going to be displayed anymore. This implementation proved to have the best results because of her general user oriented approach, showing the nodes that were most needed in the current iteration and filtering out the least used nodes from the previous iterations.

VI. Conclusion and Future Work

We have designed, implemented, and tested techniques for interactively exploring graphs in contextualized meaning. Our system enables transitions from one view to the next in an appealing manner that reduces confusion and layouts that produce transitions that are smooth and easy to follow. Using the context we are capturing the user's behaviour and using heuristics we match the best transition for the layout. D3JS is being used because it has tools that make the connection between data and graphics fairly easy. It sits right between the two, the perfect place for a library meant for data visualization. The system was evaluated and successfully tested on few datasets including the Facebook friendship dataset and a dataset containing generated relationships between students and professors.

Acknowledgments

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University".

References

[1] P. Keller and M. Keller, Visual Cues. IEEE Press, 1992.

[2] C Bennett, J Ryall, L Spalteholz, A Gooch - The Aesthetics of Graph Visualization, 2007 – Citeseer

[3] Beck, F, Burch, M., Diehl, S., Towards an Aesthetic Dimensions Framework for Dynamic Graph Visualisations, Information Visualisation, 2009 13th International Conference

[4] Blast2GO, http://www.blast2go.com/ (20.03.2014)

[5] Neo4j, http://www.neo4j.org/ (18.02.2014)

[6] D3JS, http://d3js.org/ (10.02.2014)

[7] J. Rilling, J. Wang, S. P. Mudur "MetaViz – Issues in Software Visualizing Beyond 3D", 2003.

[8] Di Battista, G., Eades, P., Tamassia, R., and Tollis, I. G., Graph Drawing: Algorithms for the Visualization of Graphs. Upper Saddle River, N. J: Prentice Hall, 1999.

[9] Wills, G. J., "NicheWorks—interactive visualization of very large graphs," Proceedings of Graph Drawing '97,1997.

Classification of images by average pixel color

Bojan Ilijoski
Faculty of Computer
Science and Engineering
Skopje, Macedonia
Email: bojan.ilijoski@finki.ukim.mk

***Abstract*—Finding the prevailing color on a image has a wide application. This paper presents a way to cluster images by their prevailing color. We are trying to find the color of an object which is presented on an image. This method should help as find out the color's name. The number of clusters is known before the start of the clustering. This means that we have already selected a palette of known colors. The main purpose of this method is to effectively find the prevailing color on the image. The idea is to detect a color on low resolution images taken by mobile phone, then get RGB or other representation of the color, and then find the closest color from the palette. This will allow us get the known name of the color which is the most similar to the color on the image. In other words, we should be able to get the most similar color from the palette to the one on the image.**

I. Introduction

The invention of images gives humanity the ability to store data in a natural form more accessible than language. The digitalization era affects images and the quantity of data that can be digitalised easily and with no or little expense in images implies the needs of image classification. Also people have easy access to cameras and can take images at any time which contributes to increase the number of images leading to a greater need for classification. A common data that images carry is color, so images can be classified by color. Classification of images by color is classical problem with many solutions and a lot of space for research and development. This paper offers an efficient algorithm for classification of image by the color dominant on it. The images used are images from a digital camera with random content, so that the results can be applied in color recognition. One great difficulty in image processing is the difference in quality of images and the difference of devices used to take photos. The task gets easier if there are palettes of colors by which we can classify. Some form of automated color distinction is of great importance when colors are close in hues and people with eyesight problems can benefit greatly. In this paper we will briefly describe the problem we are trying to solve. In the second section we will give a short description of the algorithm that we offer as a solution and in the third section we present results and efficiency of the solution.

A. The problem

The question we are trying to solve is what is the color of the subject of interest in real time and very limited processing power. As a factor in solving the problem is the size of the image. The bigger the resolution the greater the memory and the greater the processing power for processing. On the other side, small photo means small resolution that carry less data which improves the capability of the algorithm for errors. Other factor in solving the problem is the palette of colors that we use in the algorithm. Big set of colors means greater time for a photo to be classified. A good set of colors is an important input and it has to correspond the problem we are trying to solve. The set has to be as small as possible not to process redundant classification and big enough so that the answer can be solved.

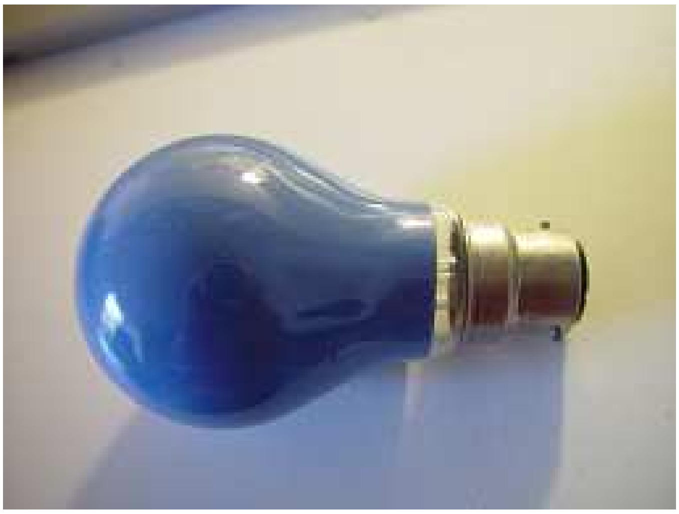


Fig. 1. What's the color of the object

B. The solution

One way to handle photos that are too big is to change them in size. We assume that the object of interest is in the center, so we process a part of the photo that is in central position. The rest of the photo gets discarded 2. This improves the overall performances of the algorithm for 90%.

A great benefit in solving this problem is the set of colors, the palette that is being offered as an input. That palette has to be predefined. The user has the ability to put the focus on certain colors. The palette is an input and it has to be created by the user. It contains a set of colors previously selected by the user. Since the algorithm's correctness is based on the quality of palette the user inputs, the problem has to be well defined. If the problem is distinction between basic colors, we input a palette consisting of basic colors. If we want a solution deeper in details we input a palette consisting of different hues. The classification of hues is possible for one color i we know the color and we input a palette of hues for that color. Smaller number of colors in a palette means better results, so if the

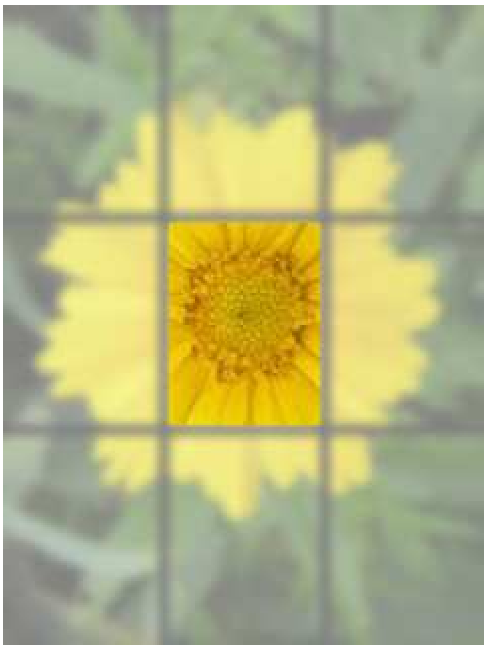


Fig. 2. Only center of the image is processed

problem is well defined a small set of colors which excludes colors that are not part of the problem can be used.

II. Background Work

There are some approaches used to solve this problem. Some of them give a comparison [1] of clustering algorithms used for color image clustering. Another approach is to make color image segmentation not based on RGB but HSI color space [2]. Some authors are using the CIE color space [3]. All this approaches are trying to represent color as close as possible to the human color perception [4]. There are different approaches not only in color representation but also in algorithms used for classification. Some of them use genetic c-Means clustering algorithm, another use competitive learning [6] or agglomerative clustering [7]. All this approaches have some advantages or disadvantages.

III. The Algorithm

In this section we will present the algorithm that solves the problem with the classification of images by their color. To begin we select the central part of the image. The rest is left out and not considered in the further processing. This enables better performances if we presume that the object of interest is in the center of the image. The second step is choosing the color palette. The colors from this palette will present the clusters. One image can be classified only by the colors suggested by the palette. The right choice of colors for the palette is crucial for the performances and the algorithm's efficiency. When the first two steps are completed we move on to the main phase which is the classification of the image by its color. In order to choose witch color the image belongs to, we use the following approach. We determine to which color a pixel belongs. Every pixel is represented in RGB color space. Also there is a possibility to use some filters for pixel smoothing before the decision. The image is classified by the color by which most of the pixels are classified. The pixels are classified by measuring the euclidean distance from each pixel to each one of the suggested colors from the palette. The colors from the palette are also represented in RGB color space. The pixel is classified by the color closest to it. Another improvement we get here is the classification of only a certain number of pixels. This leads to an improvement of the performances because it is not necessary to classify all the pixels from the image. Additionally, we can decide to use some filter for the determination of the pixel's color, which means that certain number of pixels around that one would be considered. This is practical because the human eye does not notice the color of only one pixel. Because they are really small units we always see more of them at ones. When we look at more pixels with different colors at the same time the human eye can perceive a different color. In order to improve the performances we suggest a choice of random pixels for the classification. Instead of classifying all the pixels from the image, we choose a certain number of random pixels and we classify them. Assuming that the number of pixels from the predominant color is large, this should lead to results close to the results that we would get if we classified all the pixels from the image. On the other side there would be a significant improvement in the performances because they would always have a constant number of pixels that should be classified.

The pseudo code of the algorithm is the following

```
function IMAGECLASSIFIER(image, palette)
    $image \leftarrow getImageCenter(image)$
    $pixels \leftarrow getPixels(image)$
    for all pixels do
        $color \leftarrow classify(pixel, filter)$
        $increase\_number\_of\_pixels\_classified\_in\_that\_color$
    end for
    return $color\_with\_the\_greatest\_number\_of\_pixels$
end function
```

the function $classify$ can be called with or without filter. The function $getPixels$ can return all the pixels from the image or only a part of them.

IV. Results

The testing data consisted of 110 images from 11 different colors (10 images from each color were included). We used 6 different box filters [8] and 4 ways of selection of a number of pixels included in the classification. The following color palette is used.

- black 0x000000 [0, 0, 0]
- blue 0x0000FF [0, 0, 255]
- brown 0x964B00 [150, 75, 0]
- green 0x008000 [0, 128, 0]
- orange 0xFFA500 [255, 165, 0]
- pink 0xFFCBDB [255, 192, 203]
- purple 0x800080 [128, 0, 128]
- red 0xFF0000 [255, 0, 0]
- silver 0xC0C0C0 [192, 192, 192]

1	1	1
1	1	1
1	1	1

0	1	0
1	1	1
0	1	0

Fig. 3. filter3 and filter3plus

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

0	0	1	0	0
0	1	1	1	0
1	1	1	1	1
0	1	1	1	0
0	0	1	0	0

Fig. 4. filter5 and filter5plus

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
1	1	1	1	1	1	1
0	1	1	1	1	1	0
0	0	1	1	1	0	0
0	0	0	1	0	0	0

Fig. 5. filter7 and filter7plus

- white 0xFFFFFF [255, 255, 255]
- yellow 0xFFFF00 [255, 255, 0]

The filters used are presented on the images 3, 4 and 5

On the images 6, 7, 8, 9, 10, 11 and 12 the results from the classification of the images with the use of different filters and no filters are presented. On each image are presented the results that came from the classification of all pixels on the images, on 10, 50 and 100 pixels chosen by chance for each color.

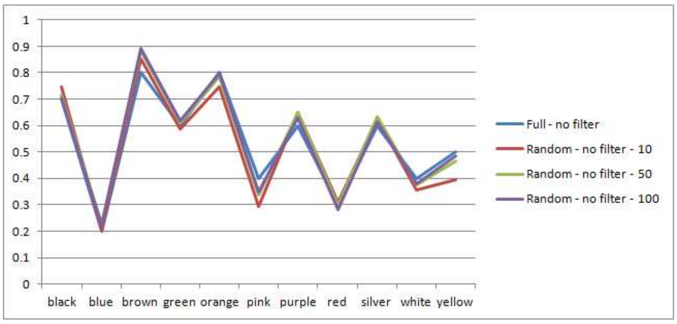


Fig. 6. precision for 4 different approaches without use of filter

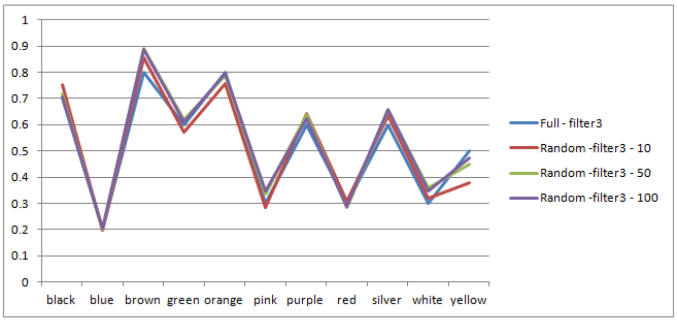


Fig. 7. precision for 4 different approaches with use of filter3

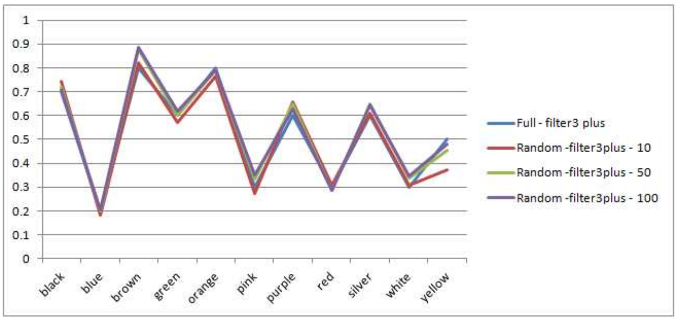


Fig. 8. precision for 4 different approaches with use of filter3plus

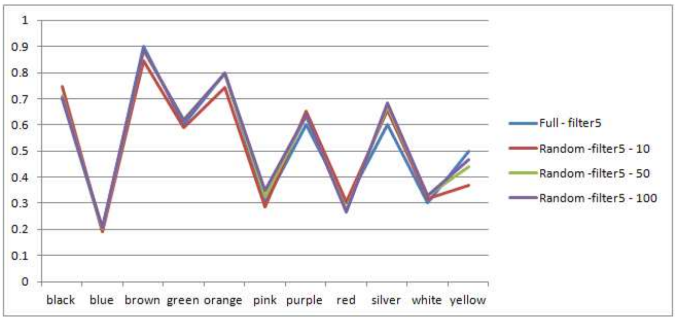


Fig. 9. precision for 4 different approaches with use of filter5

V. Conclusion

According to the results we got, we can conclude that this algorithm gave the expected ones. The performances of

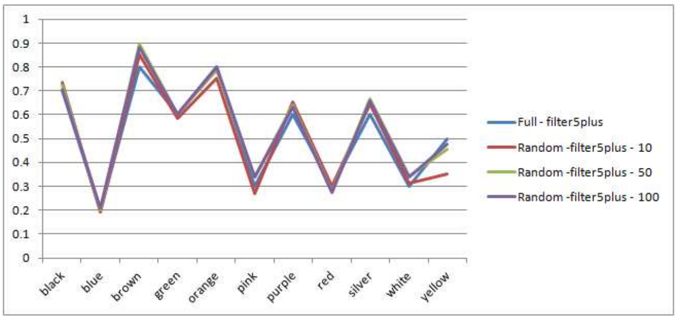


Fig. 10. precision for 4 different approaches with use of filter5plus

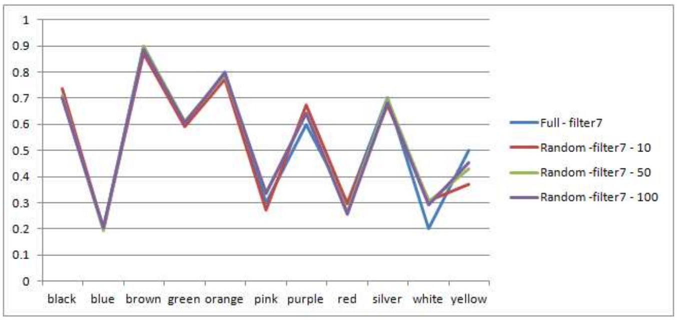


Fig. 11. precision for 4 different approaches with use of filter7

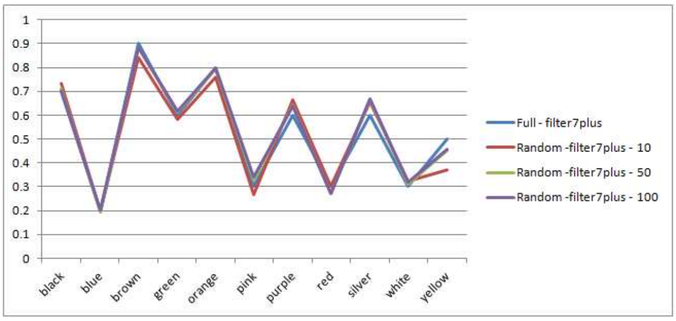


Fig. 12. precision for 4 different approaches with use of filter7plus

the image processing are significantly better if only a part of the image is used and if random pixels are chosen for the determination of the color. From the accuracy we have from the results, we can notice that the accuracy in a case when a small number of pixels are used instead of all, does not differ much, and in some clustering there is even an improvement. With the division of the image and considering only the central part of it, the number of pixels decreases for 1/9. With this, the speed needed for the processing lowers for around the same value. Even more, with considering only a certain constant number of pixels in stead of clustering all of them, the approach becomes even faster.

References

[1] P. Scheunders, A comparison of clustering algorithms applied to color image quantization, Volume 18, Issues 1113, November 1997, Pages 13791384

[2] Chi Zhang, Wang, P., A new method of color image segmentation based on intensity and hue clustering, Pattern Recognition, 2000. Proceedings. 15th International Conference on (Volume:3)

[3] Mehmet Celenk, A color clustering technique for image segmentation, Volume 52, Issue 2, November 1990, Pages 145170

[4] Yihong Gong, Proietti, G. , Faloutsos, C., Image indexing and retrieval based on human perceptual color clustering, Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on

[5] P. Scheunders, A genetic c-Means clustering algorithm applied to color image quantization, Pattern Recognition Volume 30, Issue 6, June 1997, Pages 859866

[6] Uchiyama, T, Arbib, M.A., Color image segmentation using competitive learning, Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:16 , Issue: 12)

[7] Zhigang Xiang, Gregory Joy, Color Image Quantization By Agglomerative Clustering, IEEE Computer Graphics and Applications archive Volume 14 Issue 3, May 1994 Page 44-48

[8] M. McDonnell, Box-filtering techniques Computer Graphics and Image Processing, Volume 17, Issue 1, Pages 65-70

Collection and analysis of the Macedonian on-line recipes

Ana Kuleva*, Aleksandra Bogojeska, Ilinka Ivanoska, Kire Trivadoliev, Slobodan Kalajdziski
Faculty of Computer Science and Engineering, Skopje
{aleksandra.bogojeska, ilinka.ivanoska, kire.trivodaliev, slobodan.kalajdziski}@finki.ukim.mk
*Student at Faculty of Computer Science and Engineering, Skopje
ana.kuleva@gmail.com

***Abstract*—An interesting trend emerged recently, which by using the on-line available recipes allows examination of the characteristics of the human diet, or a specific cuisine. Using this idea and the first analysis steps available, here we present the basic methods for gathering, parsing and processing recipe data, through the prism of the Macedonian cuisine. Since we also deal with specific language and alphabet, we show the workflow of the data processing and the main words parsing methods that should be taken to turn raw recipe data into cuisine specific flavour network. Additionally, the first Macedonian ingredients dictionary and the corresponding English translated dictionary are created. We also emphasise the challenges faced in the overall workflow of data preparation. The process represented here can be applied to any raw recipe data.**

***Keywords*—food pairing, online recipes, natural language processing, data extraction**

I. Introduction

The emergence of many on-line electronically available data is shifting the food science. The millions of world recipes available on-line are ideal source for detailed analysis of the today's human diet. These specific nutrition habits are mainly result of careful and long time selection of the ingredients and their combinations. But the selection of ingredients that will be combined in meal or a recipe depends of their availability in a region or territory. The main factors for cuisine distinction are the world climate and religion diversity. Different regional cuisines include different ingredients which make a cuisine recognizable. For example, the Chinese cuisine is famous for the use of rice, the Indian, curry, chilly peepers are common for the Latin American cuisine, Russians are famous for their vodka. The religion for example plays role no matter the climate or the territory, therefore the Muslims don't eat pork meat, and special dishes are cooked for Christmas holidays.

Since the ingredients are characterised with chemical compounds giving them specific flavour (as taste, odour) it is believed that the combinations of ingredients are also result of the similarity of their flavours. If two ingredients share more flavours then it is more likely that they appear in combination more often. [1], [2] This hypothesis is analysed and tested in [3], by looking into three on-line recipe repositories (www.allrecipes.com, www.epicurious.com and wwww.menupan.com). They also provide flavour compound information for the ingredients found in the recipes and by using statistics and network science the characteristic of several groups of world cuisines are obtained.

Few other publications are also dealing with the on-line recipe data. In [4] greater focus is given on the recipe recommendation algorithms by analysing not only the ingredients but the process of preparation itself, cooking methods, time, and the information stored in the user comments. Based on this information the authors propose new recipe recommendation algorithm for the users. In [5] the trait of the Chinese cuisine is analysed by looking into twenty different local Chinese cuisines and their similarities and diversities based on the climate and territorial distance. The main finding is that the territorial distance plays more important role than the climate for the similarity of the local cuisines.

Using the flavour data available in [3], here we present the first steps and the process of studying the main and specific characteristics of the Macedonian cuisine trough on-line available recipes. The recipes parsing part of the process of cuisine analysis is the most time consuming and computationally extensive and should be carried out carefully.

The paper is organized as follows: Section II presents all the steps conducted for conversion of the raw recipe data into valid ingredients with flavour information. Section III gives an overview of the resulting data of the process. Conclusion and future work information is given in Section IV

II. Methods

In this Section we are accessing the process of extraction and parsing of on-line recipe data. The workflow is presented on Figure 1. In the following, we explain all the specific steps taken to prepare the dataset for analysis.

A. Macedonian recipes web-sites

There exist several Macedonian web sites that are storing and presenting Macedonian traditional and international recipes. After looking in many sites and blogs we selected eight web sites for extraction of recipes: (kulinar.mk, moirecepti.mk, surovoivkusno.com, migusto.mk, tikves.com.mk, somelikeitraw.mk, mojatakujna.mk, chasovipogotvenje.mk). These sites were selected based on the number of recipes they store and the structure of the site itself, i.e. how easy data can be gathered from the site.

B. Flavour Network Data

At the beginning we also analysed and prepared the available data for the flavour network. There are three available

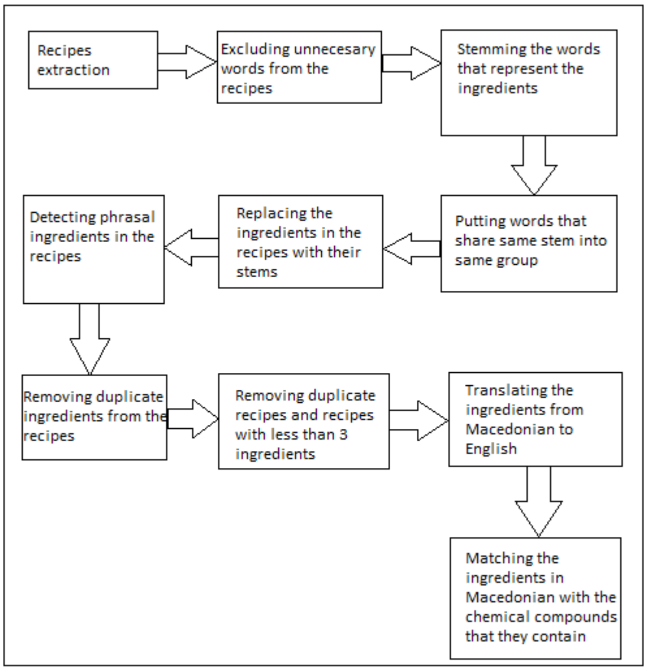
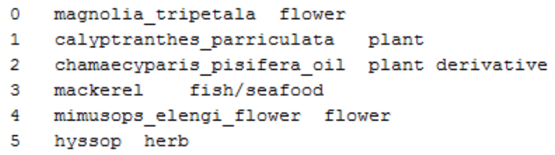

Recipes extraction
Excluding unnecesary words from the recipes
Stemming the words that represent the ingredients
Putting words that share same stem into same group
Replacing the ingredients in the recipes with their stems
Detecting phrasal ingredients in the recipes
Removing duplicate ingredients from the recipes
Removing duplicate recipes and recipes with less than 3 ingredients
Translating the ingredients from Macedonian to English
Matching the ingredients in Macedonian with the chemical compounds that they contain

Fig. 1. Workflow diagram that represents the whole process of creating corpus of the Macedonian ingredients

files. One with information for the ingredients, one for the chemical compounds and one for information about the chemical compounds of each ingredient.

Ingredients and categories data: In the ingredients database used in paper [3], a specific category is given for each ingredient. This information is stored in a file consisted of triplets “identifier ingredient category”. The ingredients are divided in 14 categories: alcoholic beverage, animal, cereal/crop, dairy, fish/seafood, flower, fruit, herb, meat, nut/seed/pulse, plant derivative, plant, spice and vegetable. Figure 2 shows part of the data file where this information is kept.

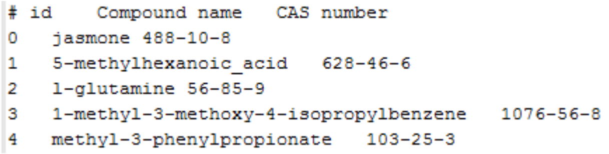


```
0   magnolia_tripetala  flower
1   calyptranthes_parriculata   plant
2   chamaecyparis_pisifera_oil  plant derivative
3   mackerel    fish/seafood
4   mimusops_elengi_flower  flower
5   hyssop  herb
```

Fig. 2. Part from the data file in which stores the ingredients and its categories

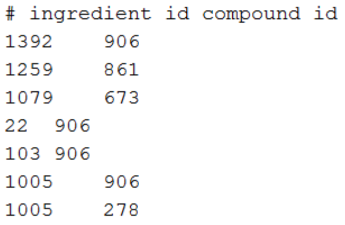
Chemical compounds data: In order the flavour network to be built, for each found ingredient there should be a set of chemical compounds that it contains. The dataset with flavour information has around 1100 chemical compounds with assigned identifiers, and they are organized in a data file as shown on Figure 3.

Ingredient-compound data: The last file stores the flavour network information. Pairs of ingredient and compound using identifiers are given for each ingredient and its flavours. The file contains 36 781 record, represented as in Figure 4.



```
# id    Compound name   CAS number
0   jasmone 488-10-8
1   5-methylhexanoic_acid   628-46-6
2   l-glutamine 56-85-9
3   1-methyl-3-methoxy-4-isopropylbenzene   1076-56-8
4   methyl-3-phenylpropionate   103-25-3
```

Fig. 3. Part of the data file which contains all used chemical compounds



```
# ingredient id compound id
1392    906
1259    861
1079    673
22  906
103 906
1005    906
1005    278
```

Fig. 4. Part of the data file which contains information for the ingredients an their chemical compounds

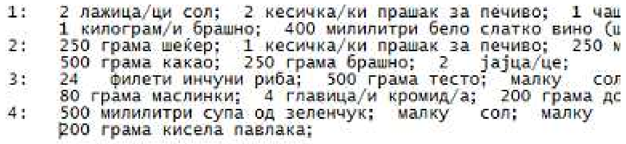
C. Crawling

For each of the selected recipes sites we developed a specific crawler in Java that extracts the important recipes information. The crawler traverses the site with a breadth first search. Pages containing the word “состојки” (ingredients) or some site specific CSS id, are further analysed for recipe extraction. If the site is well organized, one of the main criteria for selection, the ingredients are usually stored in lists, (<ul> items). If the ingredients are stored in unlabelled text boxes, paragraphs, then without any ingredients dictionary they can not be distinguished from any other words. The total information extracted for a recipe in the end includes: title and a list of ingredients, usually consisting of measurement and ingredient name.

After the process of crawling a database with around 3600 recipes was created. The extracted recipes are then subject of extensive processing using Python scripts, divided in several phases.

D. Ingredients Purification

The crawling of the web sites resulted in recipe database with very raw data including many unnecessary words. Therefore, the data was processed in few phases until a good recipe database was created. On Figure 5 the primary database of the raw recipes is showed.

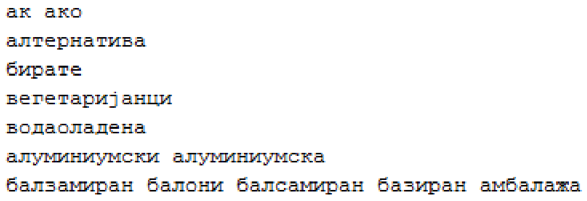


```
1:   2 лажица/ци сол;  2 кесичка/ки прашак за печиво;  1 чаш
     1 килограм/и брашно;  400 милилитри бело слатко вино (ц
2:   250 грама шеќер;  1 кесичка/ки прашак за печиво;  250 м
     500 грама какао;  250 грама брашно;  2   jajца/це;
3:   24   филети инчуни риба;  500 грама тесто;  малку   сол
     80 грама маслинки;  4 главица/и кромид/а;  200 грама дс
4:   500 милилитри супа од зеленчук;  малку   сол;  малку
     200 грама кисела павлака;
```

Fig. 5. Part of primary database of the Macedonian cuisine where each recipe is kept in separate line

1) Excluding unnecessary words: From each of the recipes extracted in the previous step, some words should be filtered out. For instance, in this dataset the ingredients amount is excess information and it should not be considered while building the database. These words include grams (грамови), millilitres (милилитри), small (малку), spoon/s (лажица/ци) and many different descriptive and stop words. But this data is not so irrelevant for other applications that use similar kind of dataset. For instance, if the application needs to make classification of some recipe and classify it as good or bad for a user suffering from specific illness this information could be of great importance. If this is the case, then how much of some ingredient is used, or how the meal is cooked, could have great meaning to the classification engine.

In order to correctly remove all unnecessary words, we store them in a special data file, called "black list". These words include the Macedonian stop words, different adjectives and words adding some information for the ingredient. The data filtering is done by manually going through the recipes and the words that are thought to be irrelevant, and writing them in the file. This process goes through few iterations such that in each iteration there are new words that are appended to the data in this file. On Figure 6 you can see one part of list of "black" words.

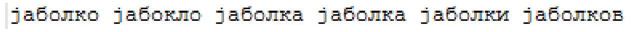


```
ак ако
алтернатива
бирате
вегетаријанци
водаоладена
алуминиумски алуминиумска
балзамиран балони балсамиран базиран амбалажа
```

Fig. 6. Part of the data file in which are stored the words that should be filtered out from the recipes

2) Words stemming: The second phase covers the process of extracting the root of the words, in our case ingredients. The stemming of the words that represent the ingredients is done such that only some of their first characters are considered to be part of their stem. If the word has length greater than 5 as a stem of the ingredient are the first 2/3 part of the word. Levenshtein distance is calculated and taken into account for finding one letter errors as: typos and mixing char sets.

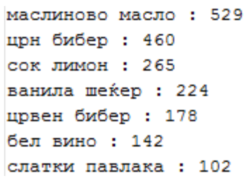
For each group of words belonging to same stemming group we keep a separate line in a file that stores these groupings. Figure 7 represents part of this data file. After all words from the recipes are grouped, the groupings are checked manually for some words that have same meaning like : maka - afion, espresso - coffee, mirudija - nane and etc. At the end each word in the recipes file is replaced with the first word of its group. For example, Figure 7 shows a case where all words should be replaced with the first word "јаболко" (apple).



```
јаболко јабокло јаболка јаболка јаболки јаболков
```

Fig. 7. Example for words that are reduced to same word (stem)

3) Phrasal ingredients: Phrasal ingredients are said to be ingredients which are represented by two or more words. Their discovering in the data is done by searching all the ingredients with length two or more and later manually going through the phrases to check their validity, since a phrase as "salt and pepper" is very common in the recipes but represents two different ingredients, compared to the "olive oil" phrase which represents one ingredient. In this process a lot of adjectives are added to the names of the ingredients because in that way they represent different ingredient from the one that is represented only by the noun form (ex. wine, red wine, white wine). This process is also done in few iterations. On Figure 8 there is a part of this data file with phrases (the numbers seen are the numbers of occurrences of the phrase in the recipes dataset).



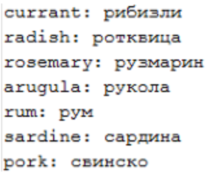
```
маслиново масло : 529
црн бибер : 460
сок лимон : 265
ванила шеќер : 224
црвен бибер : 178
бел вино : 142
слатки павлака : 102
```

Fig. 8. Part of the data file in which the phrasal ingredients are stored

Now in the recipe file are kept word phrases that are valid ingredients (olive oli, lemon juice, black pepper and etc.) and eliminated phrases that are consisted of several ingredients or can be replaced with one word (salt and pepper, onion and garlic, chicken breast, and etc.).

In this phase ingredient duplicates, which can be found in the different layers of the meal, are removed from the recipes. Duplicate recipes are removed from the recipe dataset too, since the recipes are extracted from several web sites and duplicate recipes can exist. Recipes with less then three ingredients are also remove since they don't provide any valid information.

4) Ingredients translation: In order to find the chemical compounds for the ingredients in Macedonian, they should be translated into English and mapped into the already existing ingredients in English. The only way of translating these ingredients in Macedonian is going through them manually and trying to find proper translation for them, from the database of ingredients in English from paper [3]. The problem that arises here is when the ingredient is regional and it is only present in the Macedonian cuisine, and can not be found in the existing database. As a result of this process around 350 unique translations from Macedonian to English ingredient are generated. The number of translated ingredients is even bigger but some of the ingredients in Macedonian are mapped to same ingredient in English. The diversity of the cuisines is the reason for this. Originally the dataset contains 600 ingredients in Macedonian. Figure 9 shows the mapping of the Macedonian ingredients to English, a process done completely manually.

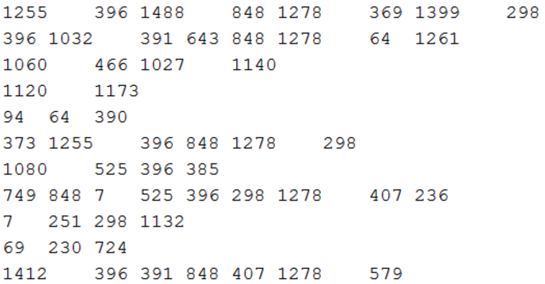


```
currant: рибизли
radish: ротквица
rosemary: рузмарин
arugula: рукола
rum: рум
sardine: сардина
pork: свинско
```

Fig. 9. Part of the data file where the mapping between the translations of ingredients in Macedonian into the ingredients in English is stored

III. Results

The extraction of the ingredients is a little bit complicated and time consuming process as could be concluded from the previous sections. After the extensive parsing of the raw recipe data as a result we got two datasets. The first dataset is purified and translated recipe data where each recipe record is consisted of ingredients identifiers. Here we have to note that the Macedonian ingredients that don't have adequate translation in the English ingredients dataset are omitted. The output database is shown on Figure 10.



```
1255    396 1488    848 1278    369 1399    298
396 1032    391 643 848 1278    64  1261
1060    466 1027    1140
1120    1173
94  64  390
373 1255    396 848 1278    298
1080    525 396 385
749 848 7   525 396 298 1278    407 236
7   251 298 1132
69  230 724
1412    396 391 848 407 1278    579
```

Fig. 10. Part of the data file with converted recipes. Ingredients are translated and replaced with identifiers

The second dataset is the flavour network defined only over the Macedonian ingredients. The ingredients from the translated dataset are assigned with flavours. This bipartite network then converted in ingredients space gives a network where each ingredient is connected with other ingredient if shares same flavours. The weight of the link is the number of same flavours they share. The file has same structure as in Figure 4.

The resulting datasets further can be used for diverse statistical measurements and analysis of the cuisine, as ingredients distribution, ingredients prevalence, most common pairs and triplets of ingredients and etc. We can test the flavour pair hypothesis and see whether the cuisine is characterised with recipes that share more similar flavours or not. The flavour network generated can be clustered and visualised using different algorithms for backbone extraction and clustering [6], [7], [8] which can provide an visual overlook of the ingredients used in the Macedonian cuisine and their flavour similarities.

IV. Conclusion

In this paper we describe the process of creating recipes and ingredients databases from the Macedonian cuisine. In order to gain this result we extracted recipes from Macedonian web sites, and as first purifying step, unnecessary, descriptive words were removed from the data. By using methods for words stemming, ingredients with same stem were classified into same group and each of the ingredients was replaced with its stem. Additionally, phrases analysis was done to identify the valid ingredient phrases. Duplicate ingredients and duplicate recipes were deleted from the dataset. As a result of these steps a database of 350 unique ingredients translated to English was created. Using the translated data chemical compounds were assigned to each ingredient. This step allowed building the Macedonian cuisine flavour network. The extracted data can be further used for diverse analysis. For instance, we can look into the frequency of each ingredient, the pairs of ingredients used more often, ingredient contribution to the cuisine, test the food pairing hypothesis and analyse the resulting flavour network by using network science methods, as clustering. The overall process described here uses natural language processing techniques and its importance is of great magnitude since is the first corpus of ingredients used in the Macedonian cuisine.

Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius University".

References

[1] H. This, *Molecular gastronomy: exploring the science of flavor*, Columbia University Press, 2005.

[2] G. A. Burdock, *Fenaroli's handbook of flavor ingredients*, CRC Press, 2004, 5th edn.

[3] Y.Y. Ahn, S.E. Ahnert, J.P. Bagrow and A.L. Barabási, "Flavor network and the principles of food pairing", *Scientific Reports*, vol. 1, pp. 196, 2011.

[4] C.Y. Teng, Y.R. Lin and L.A. Adamic, "Recipe recommendation using ingredient networks",*Proceedings of the 3rd Annual ACM Web Science Conference*, pp. 298–307, 2012.

[5] Y.X. Zhu, J. Huang, Z.K. Zhang, Q.M. Zhang, T. Zhou and Y.Y. Ahn, "Geography and similarity of regional cuisines in China",*PLOS ONE*, vol. 8, pp. e79161, 2013.

[6] M. A. Serrano, M. Boguna, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks", *Proc. Natl. Acad. Sci. U.S.A* , pp. 6483–6488, 2009.

[7] M. Rosvall, and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure", *Proc. Natl. Acad. Sci. U.S.A.* , vol. 105, pp. 1118–1123, 2008,

[8] Y.Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks", *Nature*, vol. 466.7307, pp. 761–764, 2010.

Extraction of multi-word expression from multilingual parallel corpora

Katerina Zdravkova
Faculty of Computer Science and Engineering
University Sts. Cyril and Methodius
Skopje, Macedonia
katerina.zdravkova@finki.ukim.mk

Aleksandar Petrovski
Faculty of Informatics
International Slavic University
Sveti Nikole, Macedonia
a.petrovski.sise@gmail.com

***Abstract*—Multi-word expressions (MWEs) are syntactically and semantically related groups of words. Their parsing, treatment and translation are among the most challenging natural language processing problems. The behaviour of MWEs at the lexical, syntactic and especially at the semantic level is heterogeneous. The heterogeneity increases in the internationally diverse environments. In the paper we present a system capable of extracting the candidate MWEs from various sources independently on the language the corpora were created. It consists of two interconnected phases. The initial phase ends up with a bulk of useless material, due to syntactical unimportant combinations. Syntactical filtering based on nominal and verbal phrases extracts a subset of meaningful MWEs useful for further processing. The paper concludes with summarization of the whole process and further work, including the attempt of machine translation between any of existing languages in the parallel corpora.**

Keywords— MWE extraction; multilingual parallel corpora

I. Introduction

Multi-word expressions (MWEs) are syntactically and semantically related lexical items that can be decomposed into multiple simple words [1]. They consist of two or more lexical units which have a function of a single lexical unit. For example, the adjective *artificial* and the noun *intelligence* are meaningful constituents of the expression *artificial intelligence*, which acts as an explicit term indicating a new meaning. While at the most simplistic level MWEs can be treated as sequences of simplex words in the sentence, their syntactically, and particularly semantically unexpected behaviour makes them very challenging. This behaviour is particularly evident in the idioms, which consist of related lexical units with a meaning, which is not deduced from those of the individual words [2]. Furthermore, the performance of MWEs at the lexical, syntactic and especially at the semantic level is heterogeneous. Therefore, MWEs are very important for any kind of automatic processing, predominantly when it is intended for two or more parallel languages.

In order to deal with the multi-word expressions, it is inevitable to possess at least an annotated monolingual lexicon. Macedonian language is among the languages without such a lexicon. Its creation was the crucial goal of the project presented in this paper.

The whole project was based on the multilingual resources existing in the Multext-East parallel corpora of 16 mutually sentence aligned languages [3]. In this paper, we present the present status of the extraction part, which was performed over Orwell's novel 1984. In parallel, the results of the created MWE extractor were evaluated over the Macedonian Wikipedia [4]. It will soon be extended with a new parallel multilingual edition of Jules Verne's novel "Around the World in 80 Days" [5].

The ambition to deal with MWEs in several heterogeneous multilingual environments leads to the creation of a system capable of extracting candidate MWEs from a multilingual corpus. The main prerequisite for its implementation was the existence of a compatible XML version in all the languages. The task looked rather trivial in the beginning, but it took time and lot of effort to synchronise all the resources. It is explained in more details in the second section of this paper. The third section introduces the extraction process and the problems that occurred due to the existence of smaller MWE candidates appearing in the longer phrases. The process of syntactical filtering is presented in the fourth section, together with the most valuable meaningful MWEs. The conclusion section summarizes the whole project and introduces its further development.

II. Setting up the scene

Almost all the languages existing in the Multext-East project started with the textual version of the Orwell's novel 1984 existed. Unfortunately, there was no electronic version of the novel in Macedonian. Therefore, the process of including the Macedonian language in the multilingual corpora was rather long.

The pre-processing phase started with a conversion of paper version into Microsoft Word using ABBYY FineReader, automatically checked for spelling errors and then manually corrected during machine learning of the rules for morphological analysis and synthesis of nouns, adjectives and verbs [6]. The conversion of the text into XML was done using the program UpCast, it was afterwards tokenized using the Perl program *mltokenizer*, and at the end, it was sentence aligned with the English version using the *Vanilla aligner* [7]. The alignment was manually verified and all the inconsistencies were polished [8].

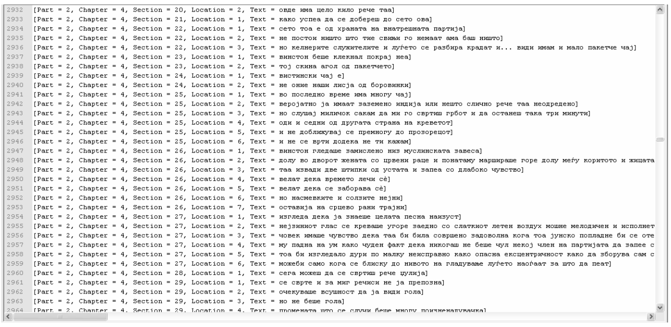
III. EXTRACTION OF CANDIDATE MWEs

The extraction of candidate multi-word expressions was done using the programming language Python in parallel on Windows and on Mac OS X [9]. The XML version was converted into a new version in which the uppercase characters were converted to lowercase, and all the punctuation marks were removed (Fig. 1.). As a result, each sentence became a continuous string of lowercase characters separated by spaces. The absence of typographic information enabled to discover the existence of exact matches, which could not exist in the Macedonian translation of the novel due to the inconsistent implementation of the delimiters [8].

The extraction process generates all the meaningful and useful MWEs. It first pulls out the blocks of words appearing at least twice. Particular attention is paid to the proper treatment of the fragments that exist in larger blocks. To achieve both goals, the process was divided into two sub-phases: separation of all the repeated blocks of words and removing of the blocks generated from longer blocks.

The first process ended up with 15463 blocks of words, which appeared 53246 times in total. The average length was 3.44 lexical units per block. The longest repeated block of words that appeared at twice in the novel consisted of 39 lexical units. The block is the following sentence: "*дури и по големите потреси и навидум неотповикливите промени секогаш се обновувал истиот модел исто како што и жироскопот секогаш се враќа во состојба на рамнотежа без состојба колку силно ќе биде турнат на една или на друга страна*". It is definitely not a useful MWE, but it contains its own blocks, some of them meaningful, such as: "*големите потреси*", "*неотповикливите промени*", "*на една или на друга страна*" or "*состојба на рамнотежа*".

Fig. 1. Part of XML version of the Macedonian version of Orwell's 1984



```
[Part = 2, Chapter = 4, Section = 20, Location = 2, Text = овде има цело кило рече таа]
[Part = 2, Chapter = 4, Section = 21, Location = 1, Text = како успеа да се добереш до сето ова]
[Part = 2, Chapter = 4, Section = 22, Location = 1, Text = сето тоа е од храната на внатрешната партија]
[Part = 2, Chapter = 4, Section = 22, Location = 2, Text = не постои ништо што тие свињи го немаат ама баш ништо]
[Part = 2, Chapter = 4, Section = 22, Location = 3, Text = но келнерите службениците и луѓето се разбира крадат и... види имам и мало пакетче чај]
[Part = 2, Chapter = 4, Section = 23, Location = 1, Text = винстон беше клекнал покрај неа]
[Part = 2, Chapter = 4, Section = 23, Location = 2, Text = тој скина агол од пакетчето]
[Part = 2, Chapter = 4, Section = 24, Location = 1, Text = вистински чај е]
[Part = 2, Chapter = 4, Section = 24, Location = 2, Text = не оние наши лисја од боровинки]
[Part = 2, Chapter = 4, Section = 25, Location = 1, Text = во последно време има многу чај]
[Part = 2, Chapter = 4, Section = 25, Location = 2, Text = веројатно ја имаат заземено индија или нешто слично рече таа неодредено]
[Part = 2, Chapter = 4, Section = 25, Location = 3, Text = но слушај миличок сакам да ми го свртиш грбот и да останеш така три минути]
[Part = 2, Chapter = 4, Section = 25, Location = 4, Text = оди и седни од другата страна на креветот]
[Part = 2, Chapter = 4, Section = 25, Location = 5, Text = и не доближувај се премногу до прозорецот]
[Part = 2, Chapter = 4, Section = 25, Location = 6, Text = и не се врти додека не ти кажам]
[Part = 2, Chapter = 4, Section = 26, Location = 1, Text = винстон гледаше замислено низ муслинската завеса]
[Part = 2, Chapter = 4, Section = 26, Location = 2, Text = долу во дворот жената со црвени раце и понатаму маршираше горе долу меѓу коритото и жицата
[Part = 2, Chapter = 4, Section = 26, Location = 3, Text = таа извади две штипки од устата и запеа со длабоко чувство]
[Part = 2, Chapter = 4, Section = 26, Location = 4, Text = велат дека времето лечи сè]
[Part = 2, Chapter = 4, Section = 26, Location = 5, Text = велат дека се заборава сè]
[Part = 2, Chapter = 4, Section = 26, Location = 6, Text = но насмевките и солзите нејни]
[Part = 2, Chapter = 4, Section = 26, Location = 7, Text = оставија на срцево рани трајни]
[Part = 2, Chapter = 4, Section = 27, Location = 1, Text = изгледа дека ја знаеше целата песна наизуст]
[Part = 2, Chapter = 4, Section = 27, Location = 2, Text = нејзиниот глас се креваше угоре заедно со слаткиот летен воздух мошне мелодичен и исполнет
[Part = 2, Chapter = 4, Section = 27, Location = 3, Text = човек имаше чувство дека таа би била совршено задоволна кога тоа јунско попладне би се оте
[Part = 2, Chapter = 4, Section = 27, Location = 4, Text = му падна на ум како чуден факт дека никогаш не беше чул некој член на партијата да запее с
[Part = 2, Chapter = 4, Section = 27, Location = 5, Text = тоа би изгледало дури по малку неисправно како опасна ексцентричност како да зборува сам с
[Part = 2, Chapter = 4, Section = 27, Location = 6, Text = можеби само кога се блиску до нивото на гладување луѓето наоѓаат за што да пеат]
[Part = 2, Chapter = 4, Section = 28, Location = 1, Text = сега можеш да се свртиш рече џулија]
[Part = 2, Chapter = 4, Section = 29, Location = 1, Text = се сврте и за миг речиси не ја препозна]
[Part = 2, Chapter = 4, Section = 29, Location = 2, Text = очекуваше всушност да ја види гола]
[Part = 2, Chapter = 4, Section = 29, Location = 3, Text = но не беше гола]
[Part = 2, Chapter = 4, Section = 29, Location = 4, Text = промената што се случи беше многу поизненадувачка]
```

This block generated two own blocks with 38 words, three blocks with 37 words etc., or in total, 741 blocks with at least two words (Table I).

TABLE I. GENERATED SMALLER MWEs

length of longer block	frequency	generated smaller blocks
39 words	2	741
24 words	2	276
20 words	2	190
17 words	2	136
12 words	2	66
In total:	10	1409

Several blocks with more than 10 words appeared exactly twice in the novel. They are actually full sentences: "*тие понатаму се делеле на разни начини носеле безброј различни имиња а нивната бројност исто како и нивниот меѓусебен однос се менувале од ера*", which consists of 24 lexical units, then the 20-words phrase "*знам дека завршуваше со еве една свеќа да ти го осветли патот еве еден џелат да ти го скине вратот*", the 17-word phrase "*можеби уште од крајот на каменото доба во светот постоеле три категории луѓе високи средни и ниски*", and at the end, the 12-word long phrase written in the Orwell's new language called Newspeak "*тајмс 03.12.1983 дневна заповед гб дуплоплуснедобро одн нелица одново*".

Whenever a smaller block generated by a longer block which already exists in the list of potential MWEs has the same frequency as the longer one, it is removed from the list of potential MWEs. This filtering process reduced the number of initial 15463 blocks of words to only 3483 potential, or so called candidate unique MWEs, resulting in a MWE corpus consisting of 22.52% of the initial blocks. They appeared 14527 times in the whole novel with an average length of 4.17 lexical units per block, which is more informative for the further treatment. Unfortunately, even with such a dramatic reduction, most of the MWEs have no value for their further processing. The list of the most frequent blocks presented in the table below proves the claim.

TABLE II. THE MOST FREQUENT CANDIDATE MWEs IN THE CORPUS

block	frequency
да се	671
да го	278
дури и	196
можеше да	193
како да	188
да биде	164
да ја	158
не се	155
што се	154
му се	143
за да	141
може да	141
не беше	137
и да	130
тоа беше	128
и со	125
не е	120
никогаш не	118
да ги	117
и на	109
и се	86
не можеше	86
да не	86
не можеше да	76
тоа што	76
му беше	63
не го	63
како и	49
сето тоа	37
тој се	37
му го	37
се случи	37
можеа да	37
како да се	37

Whenever the frequency of a smaller block exceeds the frequency of its parental block (for example, "*не можеше да*" generates both "*можеше да*" and "*не можеше*", whereas the affirmative phrase "*можеше да*" exists uniquely 117 times, while "*не можеше*" exists uniquely only 10 times), it is added to the list of potential MWEs with a unique appearance, called **CandidateUniqueMWE** list.

The **CandidateUniqueMWE** list itself still contains a raw material, because most of the blocks are not syntactically important since they don't carry any information. Such are all the 34 unique MWEs from Table 2. They usually contain the auxiliary verb *биде / сум* (*да се, не се, да биде, тоа беше, му беше*), the modal verb *може* (*можеше да, може да, можеа да, не можеше*), and many short pronominal forms that are language specific, thus completely useless in the multilingual milieu (*да го, да ја, да ги, му го*). In order to reduce the huge amount of syntactically irrelevant candidate blocks, series of useful deep structures were created, and they were incorporated in the linguistic development environment NooJ [10]. The results of the syntactical filtering are presented in the next section.

IV. SYNTACTICAL FILTERING

The starting elements for this phase were the 3483 candidate MWEs. Their frequency was at this point irrelevant, since the main goal of the phase to filter all the syntactically eligible part-of-speech combinations, called PoS sequences. To achieve the goal, all the extracted MWEs passed through a syntactical filter, which separated the insignificant MWE from potentially meaningful phrases.

Although the first grammar of the Macedonian standard language that presents the forms and their implementation was published exactly fifty years ago [11], and Macedonian online grammar created by Victor Friedman [12] exists within the joint mainframe of several Slavic and Eurasian languages, Macedonian language doesn't have a proper computational grammar. All printed grammars lack the precise deep or shallow structures, so none of there resources is useful to determine the most frequent combinations of grammatical categories capable of determining the structure of multi-word expressions [13]. Therefore, we created a small list of nominal and verbal combinations that cover the most frequent syntactical PoS combinations.

The filtering process was done with the NooJ linguistic engine [10]. The engine was selected for these main reasons:

- NooJ includes its own tools based on Finite State patterns (or, transducers) to create and maintain large-coverage lexical resources;
- Dictionaries and grammars are applied to texts in order to locate morphological, lexical and syntactic patterns of simple and compound words, making them very useful for MWEs, and
- NooJ can build complex concordances.

These Nooj features proved very successful for the creation of Macedonian computation lexicon, which comprises more than 90000 lemmas, and almost 2 million word forms [14].

The existing Macedonian lexicon already contains a small amount of more than 700 compound words, including names: *Австро-Унгарија, Ал Фатах, Охридско Езеро, Скопска Црна Гора*, short MWEs consisting of an adjective and a noun: *авторски табак, академска расправа*, dates: *две илјади четиринаесетта* etc [15].

Multext-East corpus, which generated 3483 candidate MWEs was further filtered with only two syntactic structures: the nominal and the verbal structures. Apart from them, we intend to extend the list with adjectival and propositional structures, after manual study of the written phraseological dictionaries.

A. Nominal MWEs

The list of nominal structures corresponding to nominal structures of compound words in the lexicon [14] was restricted to these PoS sequences: Adj N, Adj Adj N, Adj N N, Adj N Adj N, N Adj N, N N, N Prep N, Adj N Prep N, Adj N Adv, Adj Prep N, Adj N Prep N, Adj N Prep Adj N, and Adv N Adv N. They extracted 491 nominal phrases, most of them beneficial for the further treatment in the multilingual system. They are presented in the following table.

TABLE III. EXTRACTED NOMINAL MULTI-WORD EXPRESSIONS FROM ORWELL'S 1984

Nominal structure	Extracted examples
Adj N	*безнадежна љубов*
	човечко суштество
	ножните прсти
Adj Adj N	*друго човечко суштество*
	тивок безизразен глас
Adj N N	*мал број луѓе*
	една таблетка сахарин
Adj N Adj N	not found in CandidateUniqueMWE
N Adj N	*парче тоалетна хартија*
	црквата свети мартин
N N	*безумие безумие*
N Prep N	*член на партијата*
	вратот на микрофонот
Adj N Prep N	*будното око на полицијата*
	основната структура на општеството
	стаклен тег за хартија
	другиот крај на просторијата
Adj N Adv	*неколку минути подоцна*
Adj Prep N	*учебниците по историја*
	ораторите на партијата
	полици за книги
	министерството за изобилство
Adj N Prep N	*обичен член на партијата*
	едната рака во џебот
Adj N Prep Adj N	*дневната заповед на големиот брат*
	истакнат член на внатрешната партија
Adv N Adv N	*повеќе храна повеќе облека*

B. *Inflective forms of compound words and multi-word expressions*

The examples at the previous table show that most of the multi-word expressions contain inflective forms of adjectives and nouns. Whenever these lexical units are phraseologisms and idioms, the exact inflective form should be preserved (accepted *мати вода во аван* vs. morphologically correct, but not used: *ја мати водата во аванот*; *земи си го зборот назад*, vs. *земете си ги зборовите назад; фрли рипче*, *фати крапче* vs. *фрли го рипчето, фати го крапчето*). Multi-word expressions don't impose these limitations.

Whenever a morphological lexicon of multi-word expressions is available, NooJ is capable of recognising all their inflectional forms. For example, this is a typical lexical entry of a MWE.

кој било,PRO+FLX = KOJC0 (1)

where "*кој било*" is the compound lexical unit, PRO stands for pronoun, FLX is a reserved word and KOJC0 is the name of the paradigm which defines the inflectional behavior of the MWE. This compound word can be inflected according to the alterations of the relative *кој*, which cover three genders, the plural form, and both cases: the dative and the accusative form. They are represented with the expression:

KOJC0 = <E>/m+s+3+Cn +

<P>(*а*/f+s+3+Cn+<B>*е*/n+s+3+Cn+<B>*и*/p+3+Cn +

<B>*му*/3+Cdl + <B>*го*/3+Cal) (2)

where <E> is an empty string, <P> says "*go to the end of the previous word form*" , <B> - "*delete the last character*", m-masculine, f-feminine, n-neutral, s-singular, p-plural, 3-third person, Cn-nominative, Cdl-dative long, Cal-accusative long.

When KOJC0 is applied to "*кој било*", these 6 word forms are obtained:

кој било, m+s+3+Cn

која било, f+s+3+Cn

кое било, n+s+3+Cn

кои било, p+3+Cn

кому било, 3+Cdl

кого било, 3+Cal (3)

Orwell's novel contains very few MWEs with more that one inflected form generated by the same MWE (for example, the basic MWE *полиција на мислата* is inflected to *полицијата на мислите*, which is found several times in the larger MWEs: *член на полицијата на мислите*, *агент на полицијата на мислите*, *во рацете на полицијата на мислите*). However, many inflected forms are possible in the written and spoken language.

C. *Verbal forms*

Unlike nominal MWEs, which demonstrated a very regular and desired behaviour, the sequence of 563 smaller verbal blocks extracted the compound tenses created with the auxiliary verbs *биде / сум* (to be): *беше можно, беше неопходно, е дозволено, е невозможно* and *има* (to have): *имал право*. Therefore, verbal phrases will be explored in more details during the next stage of the system development.

V. Conclusions and Further Work

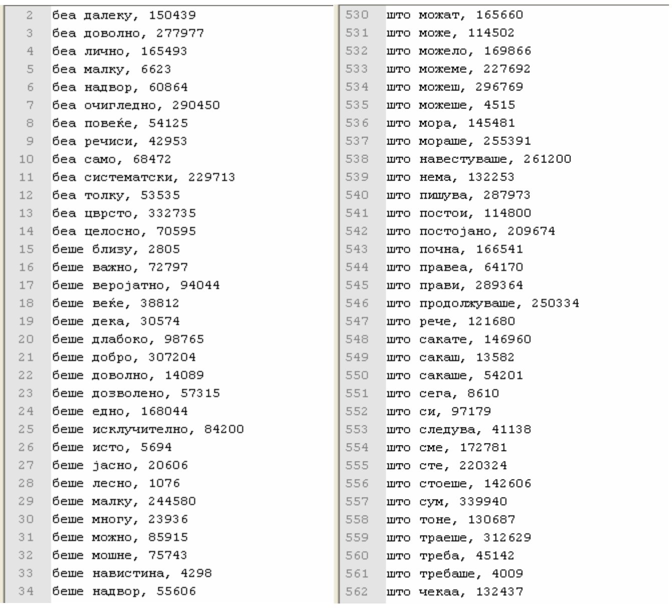
Multi-word expressions are very important lexical units, with heterogeneous behaviour at the lexical, syntactic and especially at the semantic level. Their parsing, treatment and translation are among the most challenging natural language processing problems.

In order to start their treatment and processing, it is inevitable to create a formalized dictionary of MWEs extracted from different sources.

There are many promising sources of candidate MWEs. In the project presented in this paper, we concentrated our efforts to Multext-East version 4 multilingual resources, which are based on the Orwell's novel 1984 corpora represented in 16 mutually sentence aligned languages [3]. The same approach was partially evaluated with parts of Macedonian Wikipedia, an ambitious project inspired by a similar Croatian study [16]. We were restricted by the amount of obtained material, which was in the beginning full of contents written in many languages using the Cyrillic script. It was filtered using NooJ engine, and is transferred into XML format compatible with the Multext-East format. After the extraction process, each lexical entry will comprise a MWE, its grammatical category, its inflectional class, its group and its MWE class.

The proposed approach is completely language independent. It can be implemented for any pair of bilingual parallel texts, which have been previously sentence aligned.

Fig. 2. The beginning and the end of the list of verbal MWEs



```
  2  беа далеку, 150439
  3  беа доволно, 277977
  4  беа лично, 165493
  5  беа малку, 6623
  6  беа надвор, 60864
  7  беа очигледно, 290450
  8  беа повеќе, 54125
  9  беа речиси, 42953
 10  беа само, 68472
 11  беа систематски, 229713
 12  беа толку, 53535
 13  беа цврсто, 332735
 14  беа целосно, 70595
 15  беше близу, 2805
 16  беше важно, 72797
 17  беше веројатно, 94044
 18  беше веќе, 38812
 19  беше дека, 30574
 20  беше длабоко, 98765
 21  беше добро, 307204
 22  беше доволно, 14089
 23  беше дозволено, 57315
 24  беше едно, 168044
 25  беше исклучително, 84200
 26  беше исто, 5694
 27  беше јасно, 20606
 28  беше лесно, 1076
 29  беше малку, 244580
 30  беше многу, 23936
 31  беше можно, 85915
 32  беше мошне, 75743
 33  беше навистина, 4298
 34  беше надвор, 55606
```

```
530  што можат, 165660
531  што може, 114502
532  што можело, 169866
533  што можеме, 227692
534  што можеш, 296769
535  што можеше, 4515
536  што мора, 145481
537  што мораше, 255391
538  што навестуваше, 261200
539  што нема, 132253
540  што пишува, 287973
541  што постои, 114800
542  што постојано, 209674
543  што почна, 166541
544  што правеа, 64170
545  што прави, 289364
546  што продолжуваше, 250334
547  што рече, 121680
548  што сакате, 146960
549  што сакаш, 13582
550  што сакаше, 54201
551  што сега, 8610
552  што си, 97179
553  што следува, 41138
554  што сме, 172781
555  што сте, 220324
556  што стоеше, 142606
557  што сум, 339940
558  што тоне, 130687
559  што траеше, 312629
560  што треба, 45142
561  што требаше, 4009
562  што чекаа, 132437
```

Inspired by Makoto Nagao's example based machine translation [17] and statistical machine translation used in the impressive Google Translate [18], we intend to extend our project to a system for extraction of candidate MWEs and prediction of their translations. The starting point will again be the Multext-East Version 4 corpus, with an ambition to extend it to multilingual edition of Jules Verne's novel "Around the World in 80 Days" [5].

The initial steps in this direction have also been done, and it seems that the system for machine translation of extracted candidate multi-word expression is promising because many translated Macedonian multi-word expressions to English corresponded to the extracted English multi-word expressions in the opposite direction, using the same system for extraction. In the current stage of the system, it correctly translated several smaller blocks based on statistical matching of the target sentences of aligned source sentences that carry the multi-words.

ACKNOWLEDGMENTS

Authors thank all the members of the student team involved in the implementation phase of the system for their enthusiasm and willingness to cope with all the defined tasks during the execution of various parts of the project and particularly Mario Ampov and Jovan Andonov, who contributed to the extraction of MWEs, as well as Stevica Božinovski and Mihail Petkov, who worked hard to extract and polish the contents from Macedonian Wikipedia.

This work was partially financed by the Faculty of Computer Science and Engineering at the "Sts.Cyril and Methodius University".

References:

[1] S. N. Kim, “Statistical modeling of multiword expressions”. Diss. University of Melbourne Melbourne, Australia, 2008.

[2] P. M. Matthews, “Oxford concise dictionary of linguistics”, Oxford University Press, 2005, p. 169.

[3] T. Erjavec, “MULTEXT-East: morphosyntactic resources for Central and Eastern European languages”, Language Resources and Evaluation, Volume 46, Issue 1: (2012) 131-142.

[4] Macedonian Wikipedia, http://mk.wikipedia.org/, retrieved on 30 March 2014.

[5] D. Vitas, “Multilingual Edition of Verne's Novel ‘Around the World in 80 Days’”, http://www.korpus.matf.bg.ac.rs/Verne80daysMSD/

[6] A. Ivanovska, K. Zdravkova, T. Erjavec, and S. Džeroski, “Learning rules for morphological analysis and synthesis of Macedonian nouns,adjectives and verbs” in Proceedings of 5th Slovenian and 1st international Language Technologies Conference, Jozef Stefan Institute, Ljubljana: (2006) 140-145.

[7] V. Vojnovski, S. Džeroski, and T. Erjavec, “Learning PoS tagging from a tagged Macedonian text corpus”. In Proceedings of SiKDD 2005 (Conference on Data Mining and Data Warehouses), Ljubljana, Slovenia: (2005) 199-202.

[8] M. Zlatkovska, and K. Zdravkova, “Recommendations for the Improvement of Sentence Aligners”, Proceedings of the multidisciplinary conference ‘Language and Culture Interactions via Translation and Interpreting’, 26 - 28.9. 2012, Skopje (in print)

[9] Python, http://www.python.org/, retrieved on 30 March 2014.

[10] NooJ, http://www.nooj4nlp.net/pages/nooj.html, retrieved on 30 March 2014.

[11] B. Koneski, “Grammar of Macedonian Standard Language”, Part 2: About the forms and their use, Skopje, Prosvetno delo: (1954) (in Macedonian) ”Конески, Б. „Граматика на македонскиот литературен јазик“.Дел II, За формите и нивната употреба, Скопје : "Просветно дело" : (1954)

[12] V. Friedman, “Macedonian Element Order in Declarative Sentences”, http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=3, (2001) 50-53., retrieved on 30 March 2014.

[13] R. Ivanovska, Naskova, “Development of the First LRs for Macedonian: Current Projects”, International Conference on Language Resources and Evaluation: (2006) 1837-1841.

[14] A. Petrovski, “Morphological computational dictionary - contribution to Macedonian language resources”, PhD thesis, University St. Cyril and Methodius, Faculty of Natural Sciences and Mathematics, Institute of Informatics, Skopje 2008 (in Macedonian), Петровски, А. „Морфолошки компјутерски речник - придонес кон македонските јазични ресурси“,

[15] A. Petrovski, and K. Zdravkova, “Towards a New Computational MWE Dictionary”, http://typo.uni-konstanz.de/parseme/images/Meeting/2014-03-11-Athens-meeting/PosterAbstracts/WG1-PETROVSKI-ZDRAVKOVA-1.pdf, retrieved on 30 March 2014.

[16] B. Bekavac, B., and M. Tadić, “A Generic Method for Multi Word Extraction from Wikipedia”, Proceedings of the ITI 2008 30th Int. Conf. on Information Technology Interfaces: (2008) 663-668.

[17] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle." (1984): 173-180.

[18] P. F. Brown, et al. "A statistical approach to machine translation", Computational linguistics 16.2 (1990): 79-85.

Image classification in Entomology

Martin Tashkoski and Ana Madevska Bogdanova
Ss. Cyril and Methodius University,
Faculty of Information Science and Computer Engineering,
Skopje, Macedonia
Email: tashkoskim@yahoo.com, ana.madevska.bogdanova@finki.ukim.mk

Abstract—**Digital revolution has touched most aspects of modern life, including entertainment, communication and scientific research. Digital images are just one example of this revolution. Processing of digital images is used in many areas such as applying different filters that represents photo in different ways, better quality of photos in digital cameras, extracting information from medical and microscopic images. Together image processing and machine learning offer powerful method for image classification. In this paper we present microscopic images distinction of two similar insects belonging in the family Aleyrodidae, superfamily Aleyrodoidea (whiteflies). The two insects (lat. Bemisia tabaci and lat. Trialeurodes vaporariorum) feed on plant juices and can adapt to different plants. They are very similar and they can be distinguished only in certain stage of their development (pupae stage). One of them is devastating for the plants, so it is necessary to be recognized on time. We present methodology for preprocessing and creating descriptors of the pupae images for the both insects and their classification and recognition.**

Keywords—digital images; classification; filters; microscope; entomology; svm light; weka; processing; descriptor; determination; distinction; recognition; Bemisia tabaci; Trialeurodes vaporariorum; whiteflies; insects; pupae; Vasiform orifice

I. Introduction

Agriculture has been important to humans for thousands of years. The task of feeding people is a basic rule for survival. People have always tried to improve agriculture and protect agricultural crops from various pests. People constantly learn about pests, their shape, color, development, reproduction, basic food, in order to use this knowledge in this endless fight against them.

One of this pest's problems is the problem with two whiteflies (lat. Bemisia tabaci and lat. Trialeurodes vaporariorum). Whiteflies are placed in the family Aleyrodidae and are some of the most dangerous pests for agricultural crops. They are small insects with wings and bodies all covered with a fine, powdery or flour-like white wax [1]. They do not only feed on plants, but also produce honeydew, which attracts other insects and sooty mold. They can also transmit plant viruses.

Bemisia tabaci is polyphagous pest, feeding on an estimate 600 plant species. Since the early 1980s it has caused escalating problems to both field and protected agricultural crops and ornamental plants [2]. Trialeurodes vaporariorum is also pest but its impact on the crops is less dangerous.

In Macedonia, these whiteflies can be found in the southern parts and their occurrence has been observed 5 – 6 years ago. It is assumed that they are transferred from Greece by importing various plants [3]. It is very important to stop the transfer of plants with similar pests at the border, so it would be of great benefit the construction of an intelligent system for automatic pest recognition and classification. It would be especially important the recognition of Bemisia tabaci, because it is considered more dangerous than Trialeurodes vaporariorum. Bemisia tabaci may reduce host vigour and growth, cause chlorosis and uneven ripening, and induce physiological disorders [2].

In this paper we propose a method for solving this problem and some results gained from different tests. The method is based on image processing to obtain the necessary data, which are used in Weka and SVM light in order to do the classification with different parameters and tests.

The rest of the paper is organized as follows. In the next section II we present some of the work related to our problem. In section III we present the entomology problem and the basic idea for solving this problem and in the section IV we present the results of different tests. And in the final section V we present the conclusion and future work.

II. Related work

In this section we present some of the research related to our problem for image processing and recognizing the Bemisia tabaci and Trialeurodes vaporariorum.

The main motive for beginning this idea was a research made 5-6 years ago when the authors of [3] discovered the Bemisia tabaci and Trialeurodes vaporariorum in Macedonia. They explained the danger of these pests and guess that they were transferred from other neighboring countries.

The authors in [4] made the first step for solving this problem. They proposed an algorithm for symmetrical self – filtration that can extract the important part for recognizing the microscopic image in pupae stage. Authors in [5] gave very important insight in explaining various methods for processing microscopic images. Their work on this specific area of images is very useful, because often there are dust and particles that appears as noise in the image and extraction of some information is a difficult task.

III. THE METHODOLOGY

Although Bemisia tabaci and Trialeurodes vaporariorum are very similar, there are differences in each phase of their development. But those differences are not always an accurate indicator of their distinction, because the pests can physically change their look depending on the plant from which they feed, the environment in which they live and its temperature. According to [2], the accurate indicator for distinction is based almost entirely on the fourth larval stage or “puparial stage”.

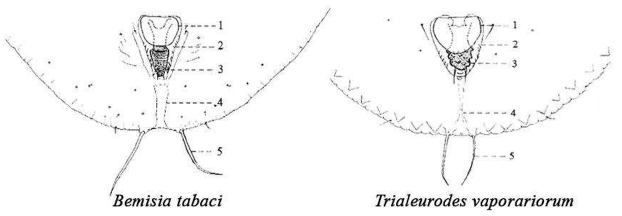


Fig. 1. Bemisia tabaci and Trialeurodes vaporariorum, closer look to “vasiform orifice”.

On Fig. 1 we can see the characteristic parts that form “vasiform orifice” and it can be noticed that the shape of the operculum (1) for both pupas is similar, but the lingula (3) of Bemisia tabaci is longer. Also the vasiform orifice (2) of Bemisia tabaci is thinner than the one of Trialeurodes vaporariorum. Shape of the lingula and vasiform orifice in the last (pupal) stage is the best indicator for recognizing the whitefly type. There are also differences in the caudal furrow (4) and caudal seta (5), but those facts are not used in this paper, because they are not clearly visible on the images that were processed.

A. Gathering images

The main difficulty for solving this problem is to find appropriate images of these two pests. The first whitefly Trialeurodes vaporariorum was found in greenhouses in the southern parts of Macedonia, on the plants cucumber and tomato. The images of the second whitefly Bemisia tabaci were found on the Internet.

There are several factors that influence the quality of the final images while capturing the whitefly Trialeurodes vaporariorum:

- *Preparation of the biological samples:*

  Different way of sample preparation can result in very different microscopic images. For example, the biological samples treated with some organic color have better quality.

- *Dust and other particles:*

  Dust and other particles are introducing noise in the images. In every biological sample there are lots of these particles. Because of the transparency of the larva, the organic substances that they were eating are appearing as noise. Because of this, the larva must be cleaned first, to see and capture the vasiform orifice.

- *Mechanical damage of the biological samples:*

  During the sample preparation, damaging the sample is very common. The larva can be easily smashed because it is very gentle.

- *Type and quality of the microscope, different zoom levels:*

  Microscope and the attached camera, zoom level, focus, lighting are also factors that affect the final images.

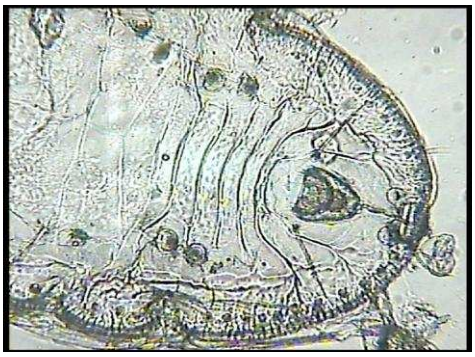


Fig. 2. Microscopic image of larva of Trialeurodes vaporariorum.

On Fig. 2 is shown microscopic image of larva of Trialeurodes vaporariorum. For the method for distinction between these two whiteflies, there are some restrictions:

- First of all, on the captured images was applied an algorithm for symmetrical self – filtration [4], and as output were got images with the characteristic part (vasiform orifice). These images are an input to the method for distinction of the whiteflies.
- The resolution of the input images with vasiform orifice is always the same.
- On the images, vasiform orifice always appeared with the thinner part down.

B. Creating desriptors

Because of the previously mentioned factors not all of the images obtained as outputs of the algorithm symmetrical self – filtration, were with good quality.

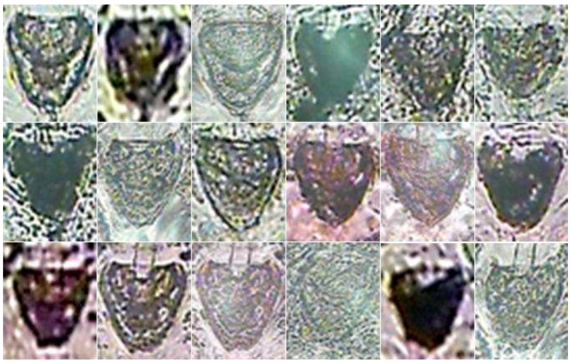


Fig. 3. Different images with “vasiform orifice” of Trialeurodes vaporariorum.

As it is shown on Fig. 3 most of the images have a lot of noise, so it is very difficult to clean them automatically. Therefore, the images used in the first tests shown in this paper, were cleaned in Photoshop with applied black and white filter.

These cleaned images with removed background and noise are used as inputs to the descriptor. The descriptor represents some parameters that describes vasiform orifice of the two whiteflies in a different way.

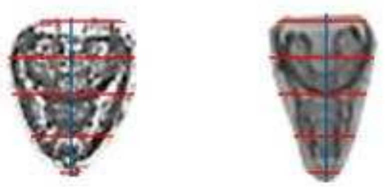


Fig. 4. Measuring the "vasiform orifice" of Trialeurodes vaporariorum and Bemisia tabaci.

IV. Tests and results

For the following tests we composed a code written in Visual Studio 2010 in C#. Cleaned images are input for the code, and the output is a file with parameters formatted properly for Weka and SVM light.

A. *File formats for Weka and SVM light*

After various tested parameters, as best descriptor was chosen the one that contains the following parameters: five widths, height, and ratio height / last width showed in Fig. 4. The resulting textual file for Weka of the two images in Fig. 4 is:

```
@relation Belokrilka
@attribute vid {tv, bt}
@attribute sirina1 real
@attribute sirina2 real
@attribute sirina3 real
@attribute sirina4 real
@attribute sirina5 real
@attribute visina real
@attribute odnosVisinaSirina real
@data
bt 19, 33, 25, 16, 6, 48, 8
tv 16, 43, 41, 34, 10, 53, 5.3
```

And the resulting textual file for SVM light is:

```
1 1:19 2:33 3:25 4:16 5:6 6:48 7:8
-1 1:16 2:43 3:41 4:34 5:10 6:53 7:5.3
```

B. *Classification in Weka and SVM light*

We have 109 images for training and testing, 48 images of Bemisia tabaci and 61 images of Trialeurodes vaporariorum. The images of Bemisia tabaci were found on Internet and we used 24 original images. We modified the original photos with some editing, in order to create bigger training set.

We made 20 tests for different classifiers in Weka and for different kernels in SVM light, and divided them in two groups: in the first 10 tests for testing we took 10 instances of Bemisia tabaci and 10 instances of Trialeurodes vaporariorum. And for the last 10 tests for training we took 40 instances of Bemisia tabaci and 40 instances of Trialeurodes vaporariorum.

C. *Results*

The first 10 tests gave the following results. For each test, we have taken 10 instances of Bemisia tabaci and 10 instances of Trialeurodes vaporariorum of the total set with 109 instances. First of all the 20 random chosen instances were training instances. After that, we made two independent testing sets, both containing 10 instances for each of the classes. For every test we performed classifications in Weka (for different classifiers) and in SVM light (for different kernels). The procedure is repeating, the chosen 20 instances are returning in the train folder, and from that folder are taken other 20 instances. In every test, for training were used 89 instances (38 instances of Bemisia tabaci and 51 instances of Trialeurodes vaporariorum).

TABLE I. Results in Weka for test file with 10 instances

Bt **Tv**	**Number of the tests**									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
trees. J48	80 %	80 %	90 %	100 %	90 %	80 %	100 %	80 %	90 %	80 %
	90 %	100 %	100 %	100 %	80 %	100 %	70 %	100 %	100 %	80 %
lazy. IBk	90 %	80 %	90 %	90 %	100 %	80 %	100 %	90 %	90 %	90 %
	100 %	100 %	100 %	80 %	90 %	100 %	80 %	100 %	100 %	100 %
bayes. Bayes Net	80 %	80 %	90 %	90 %	90 %	80 %	100 %	80 %	90 %	80 %
	100 %	90 %	100 %	90 %	100 %	100 %	70 %	100 %	100 %	80 %

Table I represents the results of the 10 tests in Weka. The white cells represent the results of Bemisia tabaci, and the grey cells present the results of the classification of Trialeurodes vaporariorum. Best results for Bemisia tabaci (recognized all test instances) we got in the 5th, and the 7th test with the classifier lazy.IBk, in the 4th and 7th with the classifier trees.J48, and in the 7th test with the classifier bayes.bayesNet. Weakest results for Bemisia tabaci (recognized 80%) we got in the 1st, 2nd, 6th, 8th, and 10th test with the classifier trees.J48, in the 2nd and 6th test with the classifier lazy.IBk, and in the 1st, 2nd, 6th, 8th, and 10th test with the classifier bayes.bayesNet. Weakest results for Trialeurodes vaporariorum (recognized 70%) were obtained in the 7th test with the classifiers trees.J48 and bayes.bayesNet.

Table II represents the results for the same previous 10 tests in SVM light. Best results for Bemisia tabaci (recognized 100%) we obtained in the 5th, and the 7th test with the kernel Radial Basis Function (for gamma = 0.01). Weakest results for Bemisia tabaci (recognized 60%) we had in the 3rd and the 6th test with the linear kernel and in the 6th test with the polynomial kernel. The results of Trialeurodes vaporariorum were very good because of the better quality of the images. Weakest results for Trialeurodes vaporariorum (recognized 90%) were obtained just with the radial basis function kernel for gamma = 0.01 in the 4th, 7th and in the 10th test.

TABLE II. RESULTS IN SVM LIGHT FOR TEST FILE WITH 10 INSTANCES

Bt Tv	Number of the tests									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Linear	80%	80%	60%	70%	80%	60%	90%	70%	70%	70%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Polynomial	80%	80%	90%	70%	90%	60%	90%	70%	80%	70%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RBF g=0.0001	80%	80%	80%	70%	90%	60%	90%	70%	80%	70%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RBF g=0.001	90%	80%	90%	90%	90%	80%	90%	70%	80%	80%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RBF g=0.01	90%	80%	90%	90%	100%	80%	100%	90%	90%	80%
	100%	100%	100%	90%	100%	100%	90%	100%	100%	90%
RBF g=0.1	20%	10%	10%	20%	20%	10%	10%	0%	10%	10%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

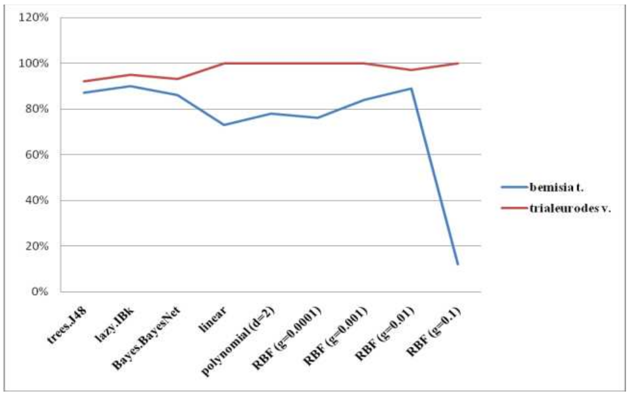


Fig. 5. Average of the results in Weka and SVM light for test file with 10 instances.

On Fig. 5 we have shown the average values of the testing instances in Weka and SVM light. According the Fig. 5, we can see that the testing with the images of Trialeurodes vaporariorum give better results.

In the second round of 10 tests, for each test, we have taken 8 instances of Bemisia tabaci and 21 instances of Trialeurodes vaporariorum of the total set with 109 instances. This was made to maintain the ratio of the instances of both classes in the train set (40 instances of Bemisia tabaci and 40 instances of Trialeurodes vaporariorum). In order to create the new testing set, we prepared 29 random chosen instances. After that, we made independent tests for each class - 8 instances of Bemisia tabaci and 21 instances of Trialeurodes vaporariorum. For every test we performed classifications in Weka (for different classifiers) and in SVM light (for different kernels).

TABLE III. RESULTS IN WEKA FOR TRAIN FILE WITH 40 INSTANCES OF BOTH CLASSES

Bt Tv	Number of the tests									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
trees.J48	87.5%	100%	100%	100%	87.5%	100%	87.5%	100%	87.5%	87.5%
	90.5%	80.9%	80.9%	90.5%	95.2%	80.9%	95.2%	95.2%	95.2%	95.2%
lazy.IBk	87.5%	87.5%	100%	100%	75%	87.5%	87.5%	100%	87.5%	75%
	95.2%	100%	90.5%	95.2%	100%	95.2%	95.2%	85.7%	100%	95.2%
bayes.BayesNet	75%	100%	100%	100%	87.5%	100%	87.5%	100%	87.5%	75%
	100%	85.7%	80.9%	90.5%	100%	95.2%	100%	95.2%	90.5%	100%

Table III represents the results for the 10 tests in Weka, where for training were used 80 instances (40 instances of Bemisia tabaci and 40 instances of Trialeurodes vaporariorum). Best results for Bemisia tabaci (recognized 100%) we obtained in the 2nd, 3rd, 4th, 6th and the 8th test with the classifier trees.J48, in the 3rd, 4th, and 8th test with the classifier lazy.IBk and in the 2nd, 3rd, 4th, 6th, and 8th test with the classifier bayes.bayesNet. Weakest results for Bemisia tabaci (recognized 75%) we had in the 5th and 10th test with the classifier lazy.IBk and in the 1st and 10th test with the classifier bayes.bayesNet. All 21 instances of Trialeurodes vaporariorum were recognized in the 2nd, 5th and 9th test with the classifier lazy.IBk, and in the 1st, 5th, 7th and 10th test with the classifier bayes.bayesNet. Weakest results for Trialeurodes vaporariorum (recognized 80.95%) were obtained in the 2nd, 3rd, and 6th test with the classifier trees.J48, and in the 3rd test with the classifier bayes.bayesNet.

Table IV represents the results of the previous tests but in SVM light. Best results for Bemisia tabaci (recognized 100%) we obtained for the 1st test with the kernel Radial Basis Function (for gamma = 0.1), in the 3rd test with all of the kernels, in the 4th test with the kernel radial basis function (for gamma = 0.01 and 0.1), in the 6th test with the kernel radial basis function (for gamma = 0.001 and 0.01), in the 8th test with the polynomial kernel and radial basis function kernel (for gamma = 0.001, 0.01 and 0.1) and in the 9th test with the radial basis function kernel (for gamma = 0.01 and 0.1). Weakest results for Bemisia tabaci (recognized 37.5%) we had in the 10th test with the radial basis function kernel (for gamma = 0.1). As the previous 10 tests the results of Trialeurodes vaporariorum were very good because of the better quality of the images. Weakest results for Trialeurodes vaporariorum (recognized 71.43%) were obtained just with the radial basis function kernel for gamma = 0.1 in the 1st and in the 4th test.

TABLE IV. RESULTS IN SVM LIGHT FOR TRAIN FILE WITH 40 INSTANCES OF BOTH CLASSES

Bt **Tv**	**Number of the tests**									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Linear	75%	62.5%	100%	87.5%	87.5%	75%	75%	87.5%	62.5%	62.5%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Polynomial	75%	62.5%	100%	87.5%	87.5%	87.5%	75%	100%	62.5%	75%
	100%	100%	100%	95.2%	100%	100%	95.2%	95.2%	100%	100%
RBF g=0.0001	75%	62.5%	100%	87.5%	87.5%	87.5%	75%	87.5%	62.5%	75%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RBF g=0.001	87.5%	75%	100%	87.5%	87.5%	100%	75%	100%	75%	75%
	95.2%	100%	95.2%	95.2%	100%	100%	95.2%	95.2%	100%	100%
RBF g=0.01	87.5%	87.5%	100%	100%	87.5%	100%	87.5%	100%	100%	75%
	95.2%	95.2%	95.2%	90.5%	100%	100%	100%	90.5%	100%	100%
RBF g=0.1	100%	50%	100%	100%	50%	75%	50%	100%	100%	37.5%
	71.4%	100%	80.9%	71.4%	100%	100%	100%	80.9%	95.2%	100%

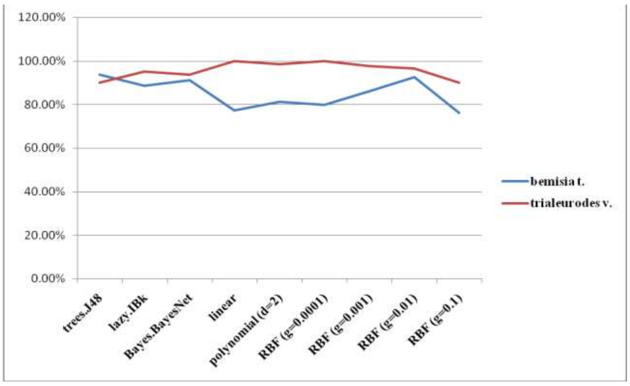


Fig. 6. Average of the results in Weka and SVM light for train file with 40 instances of both classes.

In Fig. 6 are shown the average values of the testing with 8 instances of Bemisia tabaci and 21 instances of Trialeurodes vaporariorum in Weka and SVM light. From the figure we can see that the testing with the images of Trialeurodes vaporariorum give better results, but also if we compare with Fig. 5 we can conclude that results from the second round of experiments are better.

V. CONCLUSION

In this paper we explained the method for recognition of Trialeurodes vaporariorum and Bemisia tabaci, where we included image processing and machine learning. Image processing is constantly improving in order to obtain better descriptor for the use of machine learning techniques.

We conducted two series of experiments, and showed that the even ratio of the both classes in the training set, produced better classification results. This result is illustrated by the Fig. 5 and Fig. 6

The main idea for recognizing the two whiteflies is of great benefit of the agriculture. The next step will be to make an automatic filter for cleaning the images and removing the background of the vasiform orifice. This is just a beginning of realization and improving the whole idea.

REFERENCES

[1] Gregory S. Hodges and Gregory A. Evans, "An identification guide to the whiteflies (Hemiptera: Aleyrodidae) of the southeastern United States", Taxonomic Entomologist, Florida Dept. Agriculture, Division of Plant Industry, Gainesville, FL 32614, Research Entomologist, United States Department of Agriculture, Animal plant health inspection service, Beltsville, MD, December 2005.

[2] C. Malumphy, "Protocol for the diagnosis of quarantine organism Bemisia tabaci (Gennadius)", Central Science Laboratory, Sand Hutton, York YO41 1LZ, UK.

[3] С. Банцо, Р. Русевски, „Bemisia tabaci Genad. Присуство и распространување во Република Македонија", XXXI - во традиционално советување за заштита на растенијата на Република Македонија, 21- 24 ноември 2006 година, Охрид.

[4] Марјан Киндалов, „Препознавање на слики во ентомологијата и практична имплементација со Bemisia tabaci и Trialeurodes vaporariorum" – дипломска работа, ментор: проф. Д-р Ана Мадевска Богданова, Скопје, октомври 2009.

[5] Q. Wu, Fatima A. Merchant, Kenneth R. Castleman, "Microscope image processing", ISBN: 978-0-12-372578-3, Elsevier Inc., 2008.

[6] M. H. Malais, W. J. Ravensberg, "Knowing and recognizing - The biology of glasshouse pests and their natural enemies", publisher: Koppept Biological Systems, ISBN 90 5439 126 X, 2003.

[7] Ethem Alpaydın, "Introduction to machine learning" - Second edition, The MIT press Cambridge, Massachusetts London, England, ISBN 978-0-262-01243-0, 2010.

[8] Alex Smola and S.V.N. Vishwanathan, "Introduction to machine learning", ISBN 0 521 82583 0, Cambridge University press 2008.

[9] Ian H. Witten, Eibe Frank, "Data mining" - Practical machine learning tools and techniques, second edition, ISBN: 0-12-088407-0, 2005.

[10] Gloria Menegaz (associate professor, Dept. of computer science, University of Verona (Italy)), Lessons for image processing. [Online]. Available: http://www.di.univr.it/?ent=persona&id=4526

[11] Tim Morris, "Computer vision and image processing", printed in China, ISBN 0-333-99451-5, 2004.

[12] Alan Peters, "Lectures on Image Processing", Vanderbilt University, Updated 11 March 2013.

[13] Shivam Mishra, Vaclav Hlavac and Roger Boyle, "Image Processing, Analysis, and Machine Vision", 1999, PWS Publishing, ISBN 0-534-95393-X.

[14] R. Fisher, "Digital Image Processing", 2002, Springer, ISBN 3-540-67754-2.

[15] "Morphological image processing". [Online].
Available:
https://www.cs.auckland.ac.nz/courses/compsci773s1c/lectures/ImageProcessing-html/topic4.htm

[16] Theodore Gray (Co-founder, Wolfram research), "The incredible convenience of Mathematica image processing", Wolfram blog, 2008. [Online]. Available: http://blog.wolfram.com/2008/12/01/the-incredible-convenience-of-mathematica-image-processing.

[17] Gleb V. Tcheslavski, "Morphological image processing: Basic concepts", ELEN 4304/5365 DIP, 2009.

[18] M. Ranzato, P.E. Taylor, J.M. House, R.C. Flagan, Y. LeCun, P. Perona, "Automatic recognition of biological particles in microscopic images", The Courant Institute - New York University, 719 Broadway 12th fl., New York, NY 10003 USA, California Institute of Technology 136-93, Pasadena, CA 91125 USA.

[19] CHRISTOPHER J.C. BURGES (Bell Laboratories, Lucent Technologies), "A tutorial on support vector machines for pattern

recognition", Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, 1998.

[20] Eibe Frank (Department of Computer Science, University of Waikato), "Machine learning with WEKA", New Zealand.

[21] SVMlight (official website), Support Vector Machine. [Online]. Available: http://svmlight.joachims.org

[22] Plotting and Intrepretating an ROC Curve. [Online]. Available: http://gim.unmc.edu/dxtests/roc2.htm

Session 9

Theoretical Foundations of Informatics, Security and Cryptography

One Unwanted Feature of Many Web Vulnerability Scanners

Nataša Šuteva, Dragan Anastasov, Aleksandra Mileva
Faculty of Computer Science, UGD
Štip, Macedonia
{natasa.suteva, dragan.anastasov, aleksandra.mileva}@ugd.edu.mk

Abstract - **Security experts, web developers, hackers sometimes use Web Vulnerability Scanners (WVSs) for identifying vulnerabilities in web applications. There are commercial and free/open source WVSs, and nowadays, many companies offer WVSs as services. In this paper, we test and evaluate 3 free/open source WVSs and 4 free, trial or regular editions of commercial WVSs using two versions of our one created trading web application. One version has SQL Injection and XSS vulnerabilities as critical, and the other version is free from these vulnerabilities. Results are showing that most of the scanners pollute the backend database with many garbage records using user input fields for obtaining user's opinion, comments, rating, etc., independently of the presence or absence of given critical vulnerabilities. In our experiment, garbage records were injected as comments for ads, and the magnitude of pollution goes more than 50 times the number of ads in the database in the worst case. Also, some scanners manage to find the implemented vulnerabilities without producing garbage records.**

Keywords—Web Vulnerability Scanners, backend database, garbage records

I. INTRODUCTION

Web Application Security Scanners (WASSs) or Web Vulnerability Scanners (WVSs) are a type of security software, most commonly used by website owners, security experts and hackers, to perform identification of potential vulnerabilities in the web applications, independent of the particular technology used for their implementation. They access the web applications in the same manner as user does, through the web front-end. Usually they are black-box testers, because they do not have access to the source code. Vulnerability detection mechanisms and scans differ in different WVSs, from looking at registry entries in MS Windows operating systems to see if a specific patch or update has been implemented, modifying URLs to check for sanitization issues or discover known vulnerabilities, to actually performing attacks on detecting vulnerabilities. The OWASP (Open Web Application Security Project) Top Ten 2013 vulnerability list [14] is often used as a minimum standard for website vulnerability assessment and PCI compliance according to the Payment Card Industry Data Security Standard (PCI DSS) [9], so performing web vulnerability scans is a necessity for PCI compliance. Additionally, the usefulness of WVSs comes from automatically and cost-effective conduction of security checks and production of the final report, which often includes a remedy for found vulnerability.

On the other side, WVSs are not a silver bullet, capable of detecting all of the possible vulnerabilities and attack vectors that exist. There are several reports showing that today WVSs fail to detect a significant number of vulnerabilities in test applications [2, 4, 5, 7, 8, 10, 11, 12, 13, 15].

Another big issue about WVSs is can they harm in any way tested web sites? Black box scanners have tendency to perform invasive scans, which sometimes can cause email floods, as well as publishing of garbage blog posts, garbage comments, ratings, etc [1]. Grossman [6] shares their experiences from ten years of scanning tens of thousands of real-live websites of all shapes and sizes. He gives the following 7 ways how some WVSs can harm scanned web site:

- Following "Sensitive" Hyperlinks – some web sites have hyperlinks (GET requests) that, when clicked, execute backend functionality that deletes data, cancel orders, remits payment, removes user accounts, disables functionality, and etc.
- Automatically Testing "Sensitive" Web Forms – sometimes submission of a Web form (POST request) may generate emails to customer support, execute computationally expensive backend processes, direct submitted data that will be visible to other users, and so on. This can result in spamming inboxes with thousands of emails, taking down the website due to resource load, negatively impacting the user experience of the entire user-base by showing them unexpected data, and costing the company large sums of money
- Poorly Designed Vulnerability Tests – during dynamically testing, various meta-character strings are put into input fields, URLs, POST bodies, headers, etc. Website can be harmed when it mistakes meta-characters for executable code.
- Connection Denial of Service (DoS) – sometimes scanning requires sending hundreds of requests simultaneously to the website, so this can easily exhaust a website's available connection pool and render the system unable to serve legitimate visitors.

- Session Exhaustion DoS – complete testing a website requires that vulnerability scans are run in an authenticated state. When WVS logs in hundreds of times during testing, it may consume all the website's session credential resources, and no additional legitimate users can log-in, until the session credential garbage collection is conducted.
- CPU DoS – some websites have computationally expensive hyperlinks, which during the scans may be clicked a large number of times, contrary to what was expected, and consume all of a websites available CPU resources.
- Verbose Logging and Run-Time Error – scanning can involve a large number of abnormal requests, which could raise various backend application exceptions and verbose run-time error logging. Because of this, the disk size of the logs generated and stored could be substantial.

Consequently, the vulnerability scans need to be performed with precautions, and, ideally, a replica of the live environment should be created in a test lab, so if something goes wrong, only the replica is affected. At least, before starting scans, latest backups are needed. Some automated scanners include settings for launching a non-invasive scans, but these kind of scans will only launch some very basic "security" checks against the target, such as text searches, file checks, version checks and some other basic tests, which typically do not lead to a malicious defacement of the site or web application. So, invasive scans are necessary, because if an automated WVS can break down tested website, a malicious user can do even worse.

In this paper, we try to measure the amount of generated garbage records per scan, by testing 3 free/open source WVSs and 4 free, trial or regular editions of commercial WVSs, with consideration of scanner's capability to detect several basic critical/important vulnerabilities. We want to see is it possible to detect these vulnerabilities, with performing non-invasive scans, in the sense that scanners do not leave any garbage records. Also, it was interesting to see if the pollution of database obtained by scanning, depends on the presence or absence of these vulnerabilities in the web application. After Introduction, Section II gives the basic architecture of the black box WVSs. In Section III we give a brief explanation of two versions of used testbed web application and seven WVSs with their general characteristics and input vector support, followed by used methodology, obtained results on the measured number of garbage records, and discussion. At the end, we give short concluding remarks.

II. Black Box Web Vulnerability Scanners

Generally, the core of the WVSs is made up from three main components: a crawling component, an attacker component and an analysis component.

First, the user enters at least one URL, with or without user credentials for the given web application, and then the crawling component identifies all the reachable pages in the application, and all the input points to the application. After the user sets the scanning profile, the scanner can proceed automatically or by user interaction. We used only automated mode for our experiments.

Once the crawling component finishes its job, the next components perform analysis of the discovered data, and for each web form, for each input and for each vulnerability type for which the WVS has test vectors, the component generates values that are likely to trigger a vulnerability. Then, the form content is sent to the web server as an HTTP request, and after processing the request, the server sends back a response via HTTP.

The attacker component analyzes discovered data and for each web form, for each input and for each vulnerability type for which the WVS has test vectors, the attacker module generates values that are likely to trigger a vulnerability. Then, the form content is sent to the web server using either a GET or POST request, and appropriate response is obtained from the server via HTTP.

Finally, the analysis component performs parsing and interpreting the server response. Decision if a given attack was successful is made by calculation of confidence value, by implementing attack-specific response criteria and keywords.

III. Experiments and Results

A. Testbed Web Application

We created a simple trading web application, where unregistered users can list ads, see information and description about individual ad, comment on the ad and so on. Registered users can add ads and manage ads. We created two versions of the application, a vulnerable and a safe one. The vulnerable version is affected by SQL injection (in 3 scripts), reflected and stored XSS vulnerabilities.

The web server hosting our web applications run on 64-bit Windows 8.1 Enterprise operating system. The following technologies are used: Apache server version 2.4.4, PHP version 5.4.12 and MySQL version 5.6.12.

B. Tested Web Vulnerabilities Scanners

The scanners were run on a machine with an Intel (R) Core (TM) i7-3632QM 2 x 2.20GHz CPU, 6 GB of RAM, and 64-bit Windows 8.

Table 1 lists the seven WVSs used in our study and their general characteristics. All have a graphical user interface and support for proxy mode (manual crawling). Three of them, NetSparker Community Edition, N-Stalker X Free Edition and Acunetix WVS run only on Windows, and other four can be installed on Linux and OS X also. Only N-Stalker X Free Edition, OWASP ZAP and IBM Rational AppScan can produce a report. Their input vector support is given in Table 2. Many different characteristic comparisons with older versions of these WVSs can be found on Chen's web site SecToolMarket [3].

Free NetSparker Community Edition has many features disabled, compared to its commercial version, but still you can scan and exploit SQL injection and XSS vulnerabilities without any false-positives.

TABLE 1: GENERAL CHARACTERISTICS OF THE EVALUATED SCANNERS

	NetSparker Community Edition	**N-Stalker X Free Edition**	**OWASP ZAP**	**IronWASP**	**Vega**	**Acunetix WVS**	**IBM Rational AppScan**
Company/ Creator	Mavituna Security	N-Stalker	OWASP	L. Kuppan	Sub-graph	Acunetix	IBM
Version	3.1	X-build	2.2.2	2013 beta	1.0	9	7.8
Released			Sep. 2013				
Licence/ Technology	Freeware .Net 3.5	Freeware Unknown	ASF2 Java 1.6.x	GNU .Net 2.0	EPL1 Java 1.6.x	Trial AcuSensor	Comm. Unknown
Operating System	Windows	Windows	Windows Linux OS X	Windows Linux OS X	Windows Linux OS X	Windows	Windows Linux OS X
Report	No	Yes	Yes	No	No	No	Yes
Scan Log	Yes	No	Yes	Yes	Yes	Yes	Yes

N-Stalker X Free Edition provides a restricted set of features, compared to its commercial version, and will inspect up to 100 pages within the target application. It offers a restricted version of the N - Stealth Database, web server security check, reduced analysis of web signature attacks, etc.

TABLE 2: SUPPORTING INPUT VECTORS BY THE EVALUATED SCANNERS

	NetSparker Community Edition	**N-Stalker X Free Edition**	**OWASP ZAP**	**IronWASP**	**Vega**	**Acunetix WVS**	**IBM Rational AppScan**
HTTP Query String Parameters	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HTTP Body Parameters	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HTTP Cookie Parameters	Yes	Yes		Yes		Yes	Yes
HTTP Headers	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HTTP Parameter Names				Yes			Yes
XML Element Content	Yes		Yes	Yes		Yes	Yes
XML Attributes	Yes		Yes	Yes		Yes	Yes
XML Tags							
JSON Parameters	Yes		Yes	Yes		Yes	Yes
Flash Action Message Format							Yes
Custom Input Vector				Yes			Yes
SUMMARY	**7**	**4**	**6**	**9**	**3**	**7**	**10**

OWASP Zed Attack Proxy (ZAP) is a free and open source, easy to use, integrated scanning and penetration testing tool, and it is designed to be used by people with a wide range of security experience. ZAP includes intercepting proxy, active and passive scanners, traditional and Ajax spiders, WebSocket support, fuzzing, forced browsing, port scanner, script console, etc.

IronWASP (Iron Web application Advanced Security testing Platform) is a free and open source tool, created by Lavakumar Kuppan. It offers full and semi-automated scans, JavaScript static analysis, scripting shell for Python and Ruby giving full access to the IronWASP framework, and this can be used by the pen testers to write their own fuzzers, create custom crafted request, analysis of logs, etc. Another its strength is the possibility of using different external libraries like IronPython, IronRuby, FiddleCore, etc.

Vega is a free and open source automated scanner for quick tests and an intercepting proxy for tactical inspection.

For this test we are using fully functional 14-day trial version of Acunetix WVS. This scanner uses AcuSensor Technology, and besides scanning, it offers advanced penetration testing tools.

IBM Rational AppScan, now known as IBM Security AppScan, is a family of web security testing and monitoring tools from the IBM. For our tests, we used older version of IBM Rational AppScan.

C. Methodology

In our experiments, scanners were run without logging, and only the default values for configuration parameters were used. Only N-Stalker X Free Edition was run with OWASP policy.

Backend database consists of 3 tables, with initially 3 users, 7 ads, and no comments. After every scanning we summed the number of garbage comments in the database generated by the scanner and the number of affected ads, and by deleting the comments, we prepare the database for the next scan. For every scanner we made 3 scans on the web applications.

D. Results and discussion

Table 3 shows the capabilities of tested WVSs for finding critical/important vulnerabilities. We need this to see how leaving garbage comments is connected with this capability. Only N-Stalker X Free Edition cannot find SQL vulnerability, and OWASP ZAP cannot find reflected XSS. (2/3) means that the scanner had identified only two of three vulnerable scripts.

TABLE 3: FOUNDED CRITICAL/IMPORTANT VULNERABILITIES

	SQLI	Reflected XSS	Stored XSS
NetSparker Community Edition	Yes (3/3)	Yes	Yes
N-Stalker X Free Edition		Yes	Yes
OWASP ZAP	Yes (2/3)		Yes
IronWASP	Yes (2/3)	Yes	Yes
Vega	Yes (3/3)	Yes	Yes
Acunetix WVS	Yes (2/3)	Yes	Yes
IBM Rational AppScan	Yes (2/3)	Yes	Yes

Table 4 and Table 5 give the number of garbage comments produced by the tested scanners in 3 independent scans on the safe and the vulnerable test application, respectfully.

TABLE 4: NUMBER OF GARBAGE COMMENTS FOR THE SAFE TESTBED WEB APPLICATION FOR 3 SCANS

	Number of garbage comments			Ads
	Scan 1	Scan2	Scan 3	
NetSparker Community Edition	156	160	156	All
N-Stalker X Free Edition	26	26	26	All
OWASP ZAP	61	61	61	All
IronWASP	0	0	0	-
Vega	0	0	0	-
Acunetix WVS	367	367	367	All
IBM Rational AppScan	52	52	51	All

One can see, that these numbers, ranges from 0 to 367 for the safe web application and from 0 to 180 for the vulnerable web application, and that for all scanners that produce garbage comments, all ads are affected. This means that if our database have thousands or more adds, which is the situation in reality, one scan with these scanners will produce at least the same number of the garbage comments. Two scanners, IronWASP and Vega, do not leave any garbage comments, but are capable of finding given vulnerabilities (IronWASP find 2 of 3 vulnerable scripts for SQLI). These results mean that some WVS can find tested critical/important vulnerability, without necessity to use invasive techniques. Nothing can be concluded about finding other vulnerabilities without invasive scans.

TABLE 5: NUMBER OF GARBAGE COMMENTS FOR THE VULNERABLE TESTBED WEB APPLICATION FOR 3 SCANS

	Number of garbage comments			Ads
	Scan 1	Scan 2	Scan 3	
NetSparker Community Edition	156	150	156	All
N-Stalker X Free Edition	10	10	10	All
OWASP ZAP	210	210	210	All
Iron WASP	0	0	0	-
Vega	0	0	0	-
Acunetix WVS	144	144	144	All
IBM Rational AppScan	178	178	180	All

Acunetix WVS leaves most garbage comments for the safe web application, with a magnitude of more than 50 times larger than the number of ads in the tested database. OWASP ZAP leaves most garbage comments for the vulnerable web application - 30 times larger than the number of ads in the tested database.

The experiments also show that WVSs that create garbage records, do that even when web application is free from critical/important vulnerabilities. Some WVSs, like N-Stalker X Free Edition, OWASP ZAP and Acunetix WVS produce more garbage comments for the safe web application, while IBM Rational AppScan produces more garbage comments for the vulnerable web application. First behavior is easier to understand, and can be explained that WVS stop testing the script on giving vulnerability, after it found it.

Also, some scanners, like NetSparker Community Edition and IBM Rational AppScan produce different numbers of garbage comments, but with small deviation, for scanning the same web application.

IV. CONCLUSIONS

Our experiments show that different scanners produce different numbers of garbage records in the backend database, and because of that, when we use them for scanning, web administrators need to make a backup of their database. This can protect them from spending additional time after scanning, for cleaning the database. Also, our experiments show that some scanners have capabilities of finding tested critical/important vulnerabilities, without using invasive techniques that produce garbage records. WVSs that produce garbage records, do that regardless of presence of a given vulnerability in the web application.

REFERENCES

[1] R. Abela, “A complete guide to securing a website”, Acunetix [Online]. Available: http://www.acunetix.com/websitesecurity/website-auditing-wp/

[2] J. Bau, E. Bursztein, D. Gupta and J. Mitchell, “State of the art: automated black-box web application vulnerability testing”, In Proceedings of the IEEE Symposium on Security and Privacy, May 2010.

[3] S. Chen, SecToolMarket, [Online]. Available: http://sectoolmarket.com/

[4] A. Doupe, M. Cova and G. Vigna, “Why Johnny can’t pentest: an analysis of black-box web vulnerability scanners”. In C. Kreibich, M. Jahne (Eds.) Proceedings of the 7th International conference on Detection of Intrusions and Malware, and Vulnerability Assessment - DIMVA’10, pp. 111-131, Springer Berlin Heidelberg 2010.

[5] J. Fonseca, M. Vieira, and H. Madeira, “Testing and comparing web vulnerability scanning tools for sql injection and xss attacks”, In Proceedings of the 13th IEEE Pacific Rim International Symposium. Dependable Computing (PRDC 2007), vol. 0, 2007, pp. 365–372.

[6] J. Grossman, “7 ways vulnerability scanners may harm website(s) and what to do about it”, WhiteHat Security 2012, [Online]. Available: http://blog.whitehatsec.com/7-ways-vulnerability-scanners-may-harm-websites-and-what-to-do-about-it/

[7] N. Khoury, P. Zavarsky, D. Lindskog and R. Ruhl, “Testing and assessing web vulnerability scanners for persistent SQL injection attacks”, First International Workshop on Security and Privacy Preserving in e-Societies (SeceS '11), New York, NY, USA, 2011.

[8] N. Khoury, P. Zavarsky, D. Lindskog and R. Ruhl, “An analysis of black-box web application security scanners against stored SQL Injection”, In Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT 2011) and 2011 IEEE Third International Conference on Social Computing (SOCIALCOM 2011), Boston, USA, October 2011.

[9] Payment Card Industry Security Standards Council. (PCI) Data Security Standard: Requirements and Security Assessment Procedures. October 2010. [Online]. Available: https:// www.pcisecuritystandards.org/documents/pci_dss_v2.pdf.

[10] H. Peine, “ Security test tools for web applications”. Technical Report 048.06, Fraunhofer IESE (January 2006)

[11] L. Suto, “Analyzing the effectiveness and coverage of web application security scanners”, [Online]. October 2007. Available: http://www.stratdat.com/webscan.pdf.

[12] L. Suto, “Analyzing the accuracy and time costs of web application security scanners”, [Online]. Feb 2010. Available: http://ha.ckers.org/files/Accuracy and Time Costs of Web App Scanners.pdf

[13] N. Šuteva, D. Zlatkovski, A. Mileva, “Evaluation and testing of several free/open source web vulnerability scanners”, In Proceedings of the 10th International conference on Informatics and Information Technology (CIIT 2013), 2013, pp. 221-224.

[14] Open Web Application Security Project, “OWASP Top Ten Project” [Online]. Available: http://www.owasp.org/index.php/Category: OWASP Top Ten Project.

[15] A. Wiegenstein, F. Weidemann, M. Schumacher, S. Schinzel, “Web Application Vulnerability Scanners—a Benchmark”. Technical Report, Virtual Forge GmbH (October 2006)

An Error-Detecting Code based on Linear Quasigroups

Nataša Ilievska
Faculty of Computer Science and Engeneering
Ss. Cyril and Methodius University,
Skopje, Republic of Macedonia
e-mail: natasa.ilievska@finki.ukim.mk

Danilo Gligoroski
Department of Telematics
Norwegian University of Science and Technology
Trondheim, Norway
e-mail: danilog@item.ntnu.no

***Abstract*—In this paper we consider an error-detecting code based on linear quasigroups of order 2^q defined in the following way: The input block $a_0a_1...a_{n-1}$ is extended into a block $a_0a_1...a_{n-1}d_0d_1...d_{n-1}$, where the redundant characters $d_0d_1...d_{n-1}$ are defined with $d_i = a_i * a_{i+1} * a_{i+2}$, where $*$ is a linear quasigroup operation and the operations in the indexes are per modulo n. We explain why the linear quasigroups are good for coding. Also, we give the smallest probability of undetected errors, if for coding are used quasigroups of order 8. At the end, we explain how the probability of undetected errors can be made arbitrary small.**

I. Introduction

When messages are transmitted through the noise channel, under the influence of the noise, they may be incorrectly transmitter. For this reason are used the error-detection codes, in order to detected weather the message is correctly transmitted. In this paper we consider an error-detecting code based on linear quasigroups.

Recall that quasigroup is algebraic structure $(Q, *)$ such that the equations $x * u = v$ and $u * y = v$ have unique solutions on x and y.

The quasigroup $(Q, *)$ of order 2^q is linear if there are non-singular binary matrices A and B of order $q \times q$ and a binary matrix C of order $1 \times q$, such that

$$(\forall x, y \in Q) \quad x * y = z \Leftrightarrow \boldsymbol{z} = \boldsymbol{x}A + \boldsymbol{y}B + C \qquad (1)$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are binary representations of x, y and z as vectors of order $1 \times q$ and $+$ is binary addition.

In the further work, when we say that $(Q, *)$ is a quasigroup of order 2^q, than we take $Q = \{0, 1, ..., 2^q - 1\}$.

Example 1: Suppose that we have chosen the following quasigroup:

$*$	0	1	2	3	4	5	6	7
0	5	1	0	4	3	7	6	2
1	1	5	4	0	7	3	2	6
2	7	3	2	6	1	5	4	0
3	3	7	6	2	5	1	0	4
4	0	4	5	1	6	2	3	7
5	4	0	1	5	2	6	7	3
6	2	6	7	3	4	0	1	5
7	6	2	3	7	0	4	5	1

It can be easily shown that this quasigroup is linear. Namely, since the quasigroup is of order 8, we should check if there are non-singular binary matrices A and B of order 3×3 and binary matrix C of order 1×3 such that (1) holds. The equation (1) can be represent as system of equations:

$$\begin{cases} z_1 = x_1a_{11} + x_2a_{21} + x_3a_{31} + y_1b_{11} + y_2b_{21} + y_3b_{31} + c_1 \\ z_2 = x_1a_{12} + x_2a_{22} + x_3a_{32} + y_1b_{12} + y_2b_{22} + y_3b_{32} + c_2 \\ z_3 = x_1a_{13} + x_2a_{23} + x_3a_{33} + y_1b_{13} + y_2b_{23} + y_3b_{33} + c_3 \end{cases} \qquad (2)$$

where $A = [a_{ij}]_{3\times3}$, $B = [b_{ij}]_{3\times3}$, $C = [c_i]_{1\times3}$, $\boldsymbol{x} = [x_i]_{1\times3}$ and $\boldsymbol{y} = [y_i]_{1\times3}$.

After solving the system (2) for all possible values for x and y, we get:

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

and the binary matrix C:

$$C = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$$

The matrices A and B are non-singular, so the given quasigroup is linear.

Example 2: Conversely, from the non-singular binary matrices

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

and the binary matrix C:

$$C = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

using (1) (or (2)) can be obtained the appropriate linear quasigroup.

In order to calculate $0 * 0$, we have $\boldsymbol{x} = [0\ 0\ 0]$ and $\boldsymbol{y} = [0\ 0\ 0]$, so with calculating $\boldsymbol{z} = [0\ 0\ 0]A + [0\ 0\ 0]B + [0\ 1\ 1] = [0\ 1\ 1]$, which means $z = 3$, i.e. $0 * 0 = 3$.

To calculate $0*1$, we have $\boldsymbol{x} = [0\ 0\ 0]$ and $\boldsymbol{y} = [0\ 0\ 1]$, so from (1) we have $\boldsymbol{z} = [0\ 0\ 0]A + [0\ 0\ 1]B + [0\ 1\ 1] = [1\ 0\ 0]$, which means $z = 4$, i.e. $0 * 1 = 4$.

Continuing the process with calculating all 64 products we obtain the following quasigroup:

$*$	0	1	2	3	4	5	6	7
0	3	4	0	7	5	2	6	1
1	2	5	1	6	4	3	7	0
2	0	7	3	4	6	1	5	2
3	1	6	2	5	7	0	4	3
4	4	3	7	0	2	5	1	6
5	5	2	6	1	3	4	0	7
6	7	0	4	3	1	6	2	5
7	6	1	5	2	0	7	3	4

II. Definition of the code

In [1] is defined a model of a code based on a quasigroups. In this paper we consider one special case of that code, but now defined over a set of linear quasigroups of order 2^q. Namely, we consider the following code: Let $(Q, *)$ be a linear quasigroup of order 2^q. First, each message is separated into blocks with some length n. Than, each input block $a_0a_1...a_{n-1}$, where $a_i \in Q$, is extended into a block $a_0a_1...a_{n-1}d_0d_1...d_{n-1}$. The redundant characters d_i, are defined with the following equation:

$$d_i = d_i * d_{i+1} * d_{i+2},\ i \in \{0, 1, ..., n-1\} \tag{3}$$

In the definition of this code, all operations in the indexes are per modulo n. This means that:

$$\begin{array}{lcl} d_0 & = & a_0 * a_1 * a_2, \\ d_1 & = & a_1 * a_2 * a_3, \\ d_2 & = & a_2 * a_3 * a_4, \\ . & . & . \\ . & . & . \\ . & . & . \\ d_{n-2} & = & a_{n-2} * a_{n-1} * a_0, \\ d_{n-1} & = & a_{n-1} * a_0 * a_1 \end{array}$$

Finally, the extended block $a_0a_1...a_{n-1}d_0d_1...d_{n-1}$, turned in a binary form is transmitted through the binary symmetric channel with the probability of bit error p, $(0 < p < 1/2)$. Recall that a binary symmetric channel is a channel in which inputs and outputs are zeros and ones. Under the influence of the noises, zero can be transmitted into one, or vice-versa, one in zero with probability p, while zero is transmitted into zero and one into one with probability $1 - p$, (Figure 1).

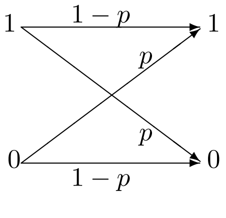


Fig. 1. Binary symmetric channel

When the receiver receives the block, it checks whether the equations (3) are satisfied for the received block. Namely, let a_i be transmitted into a_i' and d_i into d_i'. These means that the output message is $a_0'a_1'...a_{n-1}'d_0'd_1'...d_{n-1}'$. Now, the receiver checks if the following equations are satisfied:

$$\begin{array}{lcl} d_0' & = & a_0' * a_1' * a_2', \\ d_1' & = & a_1' * a_2' * a_3', \\ d_2' & = & a_2' * a_3' * a_4', \\ . & . & . \\ . & . & . \\ . & . & . \\ d_{n-2}' & = & a_{n-2}' * a_{n-1}' * a_0', \\ d_{n-1}' & = & a_{n-1}' * a_0' * a_1' \end{array} \tag{4}$$

If all equations (4) are satisfied, the receiver accepts the block. If some of the equations (4) are not satisfied, it concludes that there are errors in transmission and asks for repeated transmission of the block.

Since the redundant characters are transmitted trough the binary symmetric channel, the noises affects on them also, following that some of the redundant characters may be incorrectly transmitted. For this reason, it is possible to have undetected errors in transmission. Of course, it is good this probability to be as small as possible.

Instead directly to chose linear quasigroup of order 2^q for coding, one can chose non-singular binary matrices A and B of order $q \times q$ and binary matrix C of order $1 \times q$.

Example 3: Let say that for coding is used the linear quasigroup from Example 1.

Suppose that the input block is 245401. First, the redundant characters are calculated:

$$\begin{array}{l} d_0 = 2 * 4 * 5 = 1 * 5 = 3 \\ d_1 = 4 * 5 * 4 = 2 * 4 = 1 \\ d_2 = 5 * 4 * 0 = 2 * 0 = 7 \\ d_3 = 4 * 0 * 1 = 0 * 1 = 1 \\ d_4 = 0 * 1 * 2 = 1 * 2 = 4 \\ d_5 = 1 * 2 * 4 = 4 * 4 = 6 \end{array}$$

Now, the extended block 245401317146 turned into binary form 0101001011000**0**0001011001111001100110 is transmitted trough the binary symmetric channel. Suppose that the fourteenth character (the bolded character) is incorrectly transmitted, i.e. the receiver receives the output block 0101001011000**1**0001011001111001100110, which turned in octal form is 245421317146. In order to check whether the block is correctly transmitted, it checks whether the equations (4) are satisfied. Checking:

	calculated d_i'	received d_i'	equal?
d_0	$2 * 4 * 5 = 3$	3	yes
d_1	$4 * 5 * 4 = 1$	1	yes
d_2	$5 * 4 * 2 = 2$	7	**no**

And here it concludes that there is an error in transmission.

But, since the redundant characters are transmitted trough the channel, noises affect on them also, so it is possible for the receiver to receive the block 0101001011000**1**0001011001**0**1**0**00**0**1**1**0110, say. The bolded bits are the ones that are incorrectly transmitted. This block turned into octal form is 245421312066. Checking:

	calculated d'_i	received d'_i	equal?
d_0	$2*4*5=3$	3	yes
d_1	$4*5*4=1$	1	yes
d_2	$5*4*2=2$	2	yes
d_3	$4*2*1=0$	0	yes
d_4	$2*1*2=6$	6	yes
d_5	$1*2*4=6$	6	yes

As we can see, errors occurred, but there are not detected. So, it is possible to have undetected errors in transmission.

III. The probability of undetected errors

Naturally, when speaking for error-detecting codes it is important to now how big is the probability of undetected errors.

Since the probability of undetected errors if 4 or more characters are incorrectly transmitted and the error is not detected is inconsiderably small, when we use term probability of undetected error, we will mean on the probability that at most 3 characters of the input block are incorrectly transmitted and the error is not detected.

A. The advantage of the linear quasigroups

We showed that for the code defined with (3) the probability of undetected errors does not depend on the distribution of the characters in the input message and from the matrix C. This is one of the most important advantages of using the linear quasigroups for coding. It allow as to compute the probability of undetected errors for quasigroups of higher order, which is impossible if for coding is not used linear quasigroup.

Namely, in order to find the probability of undetected errors, we must find the class of quasigroups for which the probability of undetected errors is independent from the distribution of the characters in the input message. In the previous work with the error-detecting codes based on quasigroups ([1], [2]) the quasigroups were filtered in multiple steps, in order to find this class of quasigroups. The process of filtration was applied on the set of quasigroups of order 4. Although the filtering process is relatively slow, it can be applied on the set of quasigroups of order 4, since there are only 576 quasigroups of order 4. But it is inapplicable on the sets of quasigroups of higher order, since their number grows very rapidly with increasing n. According to [8] and [9], the number of quasigroups of order n is given with the following formula:

$$Q(n) = n!(n-1)!T(n) \qquad (5)$$

where $T(n)$ is given in Table I.

n	$T(n)$
2	1
3	1
4	4
5	56
6	9048
7	16942080
8	535281401585
9	377597570964258816
10	7580721483160132811489280
11	5363 93777 32773 71298 11967 35407 71840

TABLE I. The number of reduced quasigroups of order n

n	$T(n)$
11	$5.36 \cdot 10^{33}$
12	$1.62 \cdot 10^{44}$
13	$2.51 \cdot 10^{56}$
14	$2.33 \cdot 10^{70}$
15	$1.50 \cdot 10^{86}$

TABLE II. The aproximate number of reduced quasigroups of order n

For larger values of n, the approximate number of reduced quasigroups od order n, $T(n)$ is given in the Table II.

As we can see from the tables, the number of quasigroups rapidly grows with n. Since our code is defined over the set of linear quasigroups of order 2^n, of interest to us is the number of the quasigroups of order 2^n, $Q(2^n)$. Also from [8], the range in which are these numbers is given in the Table III.

So, for these huge number of quasigroups of relatively higher order, it is impossible to apply the method of filtration in order to get the quasigroups for which the probability of undetected errors is independent from the distribution of the characters in the input message and than to calculate this probability for all of them. Following, it is impossible to find the smallest probability of undetected errors and the best class of quasigroups for coding, i.e. the quasigroups that have the smallest probability of undetected errors. On the other hand, we showed that if linear quasigroups are used for coding, the probability of undetected errors is independent from the distribution of the characters in the input message, so there is no need of filtration. This allow as to compute the probability of undetected errors if for coding are used linear quasigroups

$$\begin{array}{ccccc}
0.101 \cdot 10^{119} & \leq & Q(2^4) & \leq & 0.689 \cdot 10^{138} \\
0.414 \cdot 10^{726} & \leq & Q(2^5) & \leq & 0.985 \cdot 10^{784} \\
0.133 \cdot 10^{4008} & \leq & Q(2^6) & \leq & 0.176 \cdot 10^{4169} \\
0.337 \cdot 10^{20666} & \leq & Q(2^7) & \leq & 0.164 \cdot 10^{21091} \\
0.304 \cdot 10^{101724} & \leq & Q(2^8) & \leq & 0.753 \cdot 10^{102805}
\end{array}$$

TABLE III. The range of the number of quasigroups of order 2^n, $n = 4, 5, 6, 7$

n	$BM(n)$
2	6
3	168
4	20160
5	9999360
6	20158709760
7	163849992929280
8	5348063769211699200

TABLE IV. THE NUMBER OF NON-SINGULAR BINARY MATRICES OF ORDER $n \times n$

n	$LQ(2^n)$
2	144
3	225792
4	6502809600
5	3199590413107200
6	26007909068026832486400
7	3436392983414413567370408755200
8	7322057236366040833007970854727843840000

TABLE V. THE NUMBER OF LINEAR QUASIGROUPS OF ORDER 2^n

of higher order and to find the best class of quasigroups for coding.

The number of binary matrices $BM(n)$ of order $n \times n$ is given in the table IV.

Consequently, the number of linear quasigroups $LQ(2^n)$ of order 2^n in given in the table V.

Further, since we showed that the probability of undetected errors does not depend from the matrix C, there is not need to calculate the probability of undetected errors for all linear quasigroups of order 2^n, but only for the all pairs of non-singular binary matrices A and B of order $n \times n$. The number of quasigroups of order 2^n for which the probability of undetected errors should be calculated, $LQC(2^n)$ is given in Table VI.

Example 4: According to formula (5) and Table I, the number of quasigroups of order 8 is 108776032404012288000.

n	$LQC(2^n)$
2	36
3	28224
4	406425600
5	99987200409600
6	406373579187919257600
7	26846820182925105995081318400
8	28601786079554847003937386151280640000

TABLE VI. THE NUMBER OF LINEAR QUASIGROUPS OF ORDER 2^n FOR WHICH THE PROBABILITY OF UNDETECTED ERRORS SHOULD BE CALCULATED

On the other hand, the number of non-singular binary matrices of order 3 is only 168, i.e. there are $168^2 \cdot 2^3 = 28224 \cdot 8 = 225792$ linear quasigroups of order 8. But since the probability of undetected error does not depend on matrix C, we should calculate the probability of undetected errors only for $168^2 = 28224$ linear quasigroups of order 8. So, instead to work with the huge number of quasigroups of order 8 and filter them in order to get the ones for which the probability of undetected errors is independent from the distribution of the characters in the input message (which is practically impossible), we work directly with this relatively small number of quasigroups that satisfy the required condition.

B. The probability of undetected errors for quasigroups of order 8

Using the above explained properties of the linear quasigroups and some combinatorics we found the smallest probability of undetected errors if for coding are used linear quasigroups of order 8. The probability of undetected errors $f(n,p)$ is a function of the block length n and the probability of bit-error of the binary-simmetric channel p. The graph of this function is given in a Figure 2.

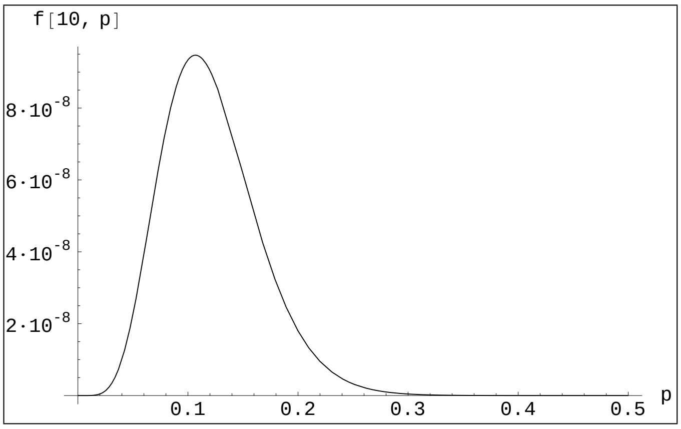


Fig. 2. The best probability functions for block length $n = 10$

The graph of the function of the probability of undetected errors for different values of n is given in a Figure 3.

From the Figure 3 we can see that this code has the nice property, the series of maximums of its probability of undetected errors to tend to zero when block length tends to infinity. Using this property, we can control the probability of undetected errors. If we want the probability of undetected errors to be smaller than ε, we chose the block length to be n, where n is chosen such that the maximum of the function $f(n,p)$ to be smaller than ε.

There are only 6 pairs of all 28224 pairs of non-singular binary matrices of order 3 for which this minimum is achieved (or only 48 linear matrices of all 225952 linear matrices of order 8).

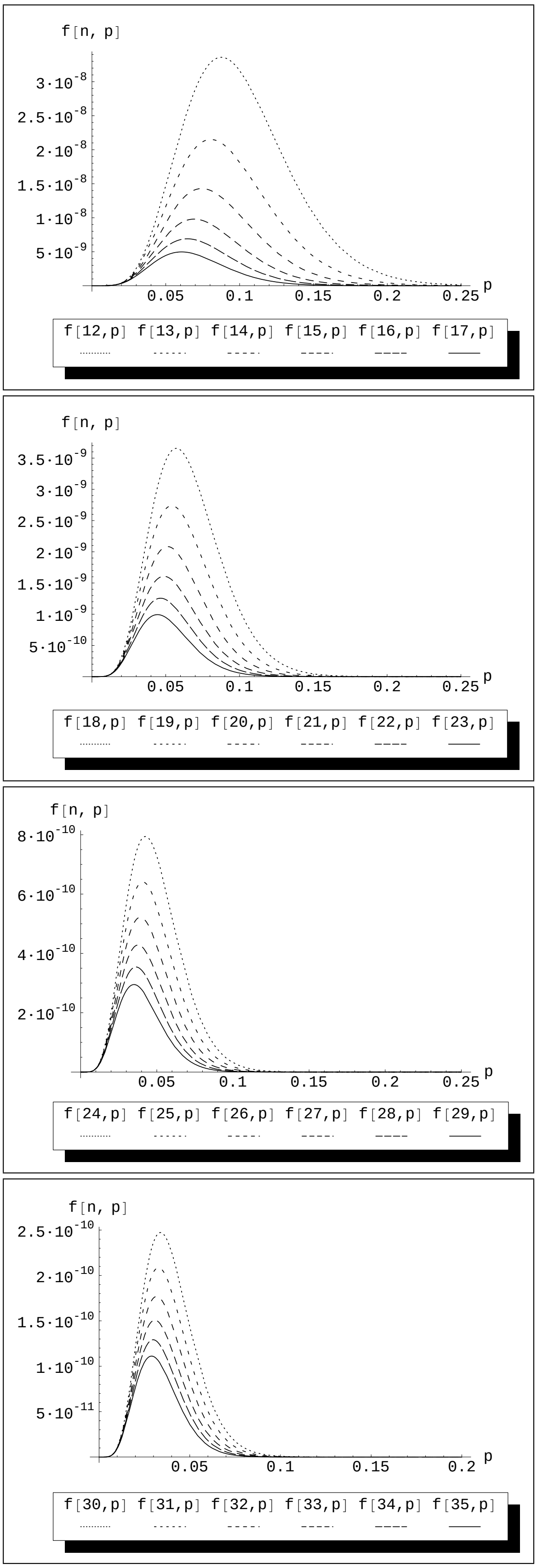


Fig. 3. The probability functions of undetected errors for the best class of linear quasigroups of order 8. Note that the scaling on y axis is different and is with different resolution.

IV. Conclusion

In this paper we consider an error-detecting code based on linear quasigroups. We concluded that the liner quasigroups are good for coding. Also, we give the smallest probability of undetected errors if for coding are used quasigroups of order 8 and the number of quasigroups that have smallest probability of undetected errors, i.e. the number of quasigroups of order 8 which are the best for coding. At the end, we explain how the probability of undetected errors can be made arbitrary small.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the ”Ss.Cyril and Methodius” University.

References

[1] V. Bakeva, N. Ilievska, *A probabilistic model of error-detecting codes based on quasigroups*, Quasigroups and Related Systems, 17, pp. 135-148, 2009.

[2] N. Ilievska, V. Bakeva, *A Model of error-detecting codes based on quasigroups of order 4*, Proc. Sixth International Confer. Informatics and Information Technology, Bitola, Republic of Macedonia pp. 7-11, 2008.

[3] Y. Chen, M. Niemenmma, A.J. Han Vinck, D. Gligoroski, *On the Error Detection Capability of One Check Digit*, IEEE Transactions on Information theory pp. 261-270, 2014.

[4] The On-line Encyclopedia of Integer Sequences, http://oeis.org/search?q=nonsingular+binary+matrices&sort=&language=&go=Search

[5] D. Gligoroski, V. Dimitrova, S. Markovski, *Quasigroups as Boolean functons, their equation systems and Gröbner bases*, short-note for RISC Book Series, Springer, ”Groebner, Coding, and Cryptography”, Ed. T.Mora, L.Perret, S.Sakata, M.Sala, and C.Traverso, pp. 415-420, 2009.

[6] S. Markovski, V. Bakeva, *On Error-detecting codes based on quasigroup operation*, Proc. Fourth International Confer. Informatics and Information Technology, Bitola, Republic of Macedonia, pp. 400-405, 2003.

[7] S. Markovski, V. Bakeva *Error-detecting codes with cyclically defined redundancy*, Proc. Third Congress of Math. of Macedonia, pp. 485-492, 2005.

[8] C. Koscielny, *Generating quasigroups for cryptocraphic applications*, Int. J. Appl.Math.Comput.Sci., 12, pp. 559-569, 2002.

[9] B. D. McKay, I. M. Wanless, *On the number of Latin squares*, Annals of Combinatorics, 9, pp. 335-344, 2005.

Applying Error-Correcting Codes Based on Quasigroups for Image Coding

Aleksandra Popovska-Mitrovikj, Daniela Mechkaroska* and Verica Bakeva
Faculty of Computer Science and Engineering
UKIM
Skopje, Macedonia
{aleksandra.popovska.mitrovikj, verica.bakeva}@finki.ukim.mk, daniela-mec@hotmail.com

Abstract— **Error-correcting random codes based on quasigroup transformations are proposed elsewhere by Danilo Gligoroski, Smile Markovski and Ljupco Kocarev. Cut-decoding algorithm is a modification of random codes based on quasigroups proposed elsewhere, also. The performances of these codes are investigated in several papers. In this paper we investigate performances of these codes for coding/decoding images. We present several experimental results obtained for different values of bit-error probability of binary-symmetric channel. These results are compared with suitable results obtained with Reed-Solomon codes.**

Keywords—error-correcting codes; image; quasigroup; quasigroup transformation; Reed-Solomon code; bit-error and packet-error probability.

I. INTRODUCTION

The Random Codes Based on Quasigroups (RCBQ) are defined in [1]. In papers [2] and [3], authors investigated the influence of the code parameters on the performances of these codes and compared these codes with Reed-Muller and Reed-Solomon codes. In [4], authors proposed some modifications of the standard coding/decoding algorithms in order to improve the code performances. Here, we investigate performances of these codes for transmitting images through binary-symmetric channel. For that aim we made experiments using standard coding/decoding algorithms and cut-decoding algorithm (with backtracking) proposed in [4]. For comparison we made experiments with Reed-Solomon codes.

The RCBQ are designed using algorithms for encryption/decryption from the implementation of TASC (Totally Asynchronous Stream Ciphers) by quasigroup string transformations [5]. These cryptographic algorithms use the alphabet Q and a quasigroup operation * on Q together with its parastrophe \. The notions of quasigroups and quasigroup string transformations are given in the previous paper for these codes ([2], [3]). Here, we are using the same terminology and notations as there. In the next section we will briefly repeat only the coding/decoding algorithms of RCBQ.

II. DESCRIPTION OF RCBQ

A. Description of coding with standard algorithm and cut-decoding algorithm

Let $M = m_1m_2\ldots m_l$ be a block of $N_{block}= la$ bits where $m_i \in Q$ and Q is an alphabet of a-bit symbols. First, we add a redundancy as zero symbols and produce message $L=L^{(1)}\ L^{(2)}\ldots L^{(s)} = L_1L_2\ldots L_m$ of N bits, where $L^{(i)}$ are sub-blocks of r symbols from Q and $L_i \in Q$ (so, $sr = m$). After erasing the redundant zeros from each $L^{(i)}$, the message L will produce the original message M. In this way we obtain an (N_{block}, N) code with rate $R= N_{block}\ /N$. The codeword is produced after applying the encryption algorithm of TASC (given in Fig. 1) on the message L. For that aim, previously, a key $k=k_1k_2\ldots k_n \in Q^n$ should be chosen. The obtained codeword of M is $C=C_1C_2\ldots C_m$, where $C_i \in Q$.

Encryption	**Decryption**
Input: Key $k = k_1k_2\ldots k_n$ and message $L = L_1L_2\ldots L_m$ **Output**: message (codeword) $C = C_1C_2\ldots C_m$	**Input**: The pair $(a_1\ a_2\ldots\ a_s,\ k_1k_2\ldots k_n)$ **Output**: The pair $(c_1\ c_2\ldots\ c_s,\ K_1K_2\ldots K_n)$
For $j = 1$ to m $X \leftarrow L_j$; $T \leftarrow 0$; For $i = 1$ to n $X \leftarrow k_i * X$; $T \leftarrow T \oplus X$; $k_i \leftarrow X$; $k_n \leftarrow T$; **Output**: $C_j \leftarrow X$	For $i = 1$ to n $K_i \leftarrow k_i$; For $j = 0$ to $s - 1$ $X, T \leftarrow a_{j+1}$; $temp \leftarrow K_n$; For $i = n$ down to 2 $X \leftarrow temp \setminus X$; $T \leftarrow T \oplus X$; $temp \leftarrow K_{i-1}$; $K_{i-1} \leftarrow X$; $X \leftarrow temp \setminus X$; $K_n \leftarrow T$; $c_{j+1} \leftarrow X$; **Output**:$(c_1c_2\ldots\ c_s,\ K_1K_2\ldots K_n)$

Fig. 1. Algorithms for encryption and decryption.

In the cut-decoding algorithm, instead of using a (N_{block}, N) code with rate R, we use together two $(N_{block}, N/2)$ codes with rate $2R$ for coding/decoding a same message of N_{block} bits. Namely, for coding we apply two times the encryption algorithm, given in Fig. 1, on the same redundant message L using different parameters (different keys or quasigroups). In this way we obtain the codeword of the message as concatenation of the two codewords of $N/2$ bits.

This research was partially supported by Faculty of Computer Science and Engineering at the University "Ss Cyril and Methodius" in Skopje.

*At the time of writing the author was a student at the FCSE

B. *Description of decoding with standard algorithm and cut-decoding algorithm*

After transmission through a noise channel (for our experiments we use binary symmetric channel), the codeword C will be received as message $D = D^{(1)}D^{(2)}...D^{(s)} = D_1D_2...D_m$, where $D^{(i)}$ are blocks of r symbols from Q and $D_i \in Q$. The decoding process consists of four steps: (*i*) procedure for generating the sets with predefined Hamming distance, (*ii*) inverse coding algorithm, (*iii*) procedure for generating decoding candidate sets and (*iv*) decoding rule.

The probability that $\leq t$ bits in $D^{(i)}$ are not correctly transmitted is $P(p;t)=\sum_{k=0}^{t}\binom{r\cdot a}{k}p^k(1-p)^{r\cdot a-k}$, where p is probability of bit-error in a binary symmetric channel. Let B_{max} be an integer such that $1–P(p;\, B_{max}) \leq q_B$ and

$$H_i = \{\alpha \mid \alpha \in Q^r, H(D^{(i)}, \alpha) \leq B_{max}\},$$

for i =1, 2, …, s, where $H(D^{(i)}, \alpha)$ is the Hamming distance between $D^{(i)}$ and α.

The decoding candidate sets S_0, S_1,…, S_s are defined iteratively. Let $S_0 = (k_1k_2…k_n;\ \lambda)$, where λ is the empty sequence. Let $S_{i\,-1}$ be defined for $i \geq 1$. Then S_i is the set of all pairs $(\delta,\ w_1w_2…w_{rai})$ obtained by using the sets $S_{i–1}$ and H_i as follows (w_j are bits). For each element $\alpha \in H_i$ and each $(\beta,\ w_1w_2…w_{ra(i-1)}) \in S_{i–1}$, we apply the inverse coding algorithm (i.e. algorithm for decryption given in Fig. 1) with input (α, β). If the output is the pair (γ, δ) and if both sequences γ and $L^{(i)}$ have the redundant zeros in the same positions, then the pair $(\delta,\ w_1w_2…w_{ra(i\text{-}1)}\ c_1c_2...c_{ra}) \equiv (\delta,\ w_1\ w_2…w_{rai})$ is an element of S_i.

The decoding of the received message D is given by the following rule: If the set S_s contains only one element $(d_1…d_n,\ w_1w_2…w_{ras})$ then $L = w_1\ w_2…w_{ras}$ is the decoded (redundant) message (then we say we have a *successful decoding*). In the case when the set S_s contains more than one element, we say that the decoding of D is unsuccessful (of type *more-candidate-error*). In the case when $S_j = \varnothing$ for some $j \in \{1, …, s\}$, the process will be stopped (we say that a *null-error* appears).

In the cut-decoding algorithm, after transmitting through a noise channel, we divide the outgoing message $D = D^{(1)}D^{(2)}...D^{(s)}$ in two messages $D^1 = D^{(1)}D^{(2)}...D^{(s/2)}$ and $D^2 = D^{(s/2+1)}D^{(s/2+2)}...D^{(s)}$ with equal lengths and we decode them in parallel with the corresponding parameters. In this method of decoding we make modification in the procedure for generating decoding candidate sets. Namely, after each iteration we obtain the decoding candidate set with elimination of all elements in the first decoding candidate set whose second part does not matches with the second part of an element in the second set, and vice versa. With this algorithm we obtained improvement in the speed of decoding and the values of PER and BER for code (72,288) using nibbles.

C. *Method for decreasing the number of null-errors*

In [2] authors proposed a method for decreasing the number of *null-errors* by backtracking. Namely, if in the i^{th} iteration of the decoding process we obtain $S_i = \varnothing$ it means that in some previous step $j < i$ we have lost the right block that had to be processed. This happened if the number of errors during transmission in some block is greater than B_{max}. Some of these errors will be eliminated if we cancel a few of iterations of the decoding process and we reprocess all of them or part of them with a larger value of B_{max}. By applying this method in cut-decoding algorithm improvements of the values of PER and BER are obtained in [3]. Therefore, we use this method for reducing *null-errors* in our experiments with cut-decoding algorithm.

III. Experiments

We made experiments with the image given in the Figure 2 and we used binary symmetric channel with bit-error probability p. In our experiments we use the following values of bit-error probability: p=0.03, p=0.06, p=0.09, p=0.12. For each value of p, we consider the image after transmission through the channel in the following cases:

- without using any error-correcting code,
- using RCBQ with the standard algorithm,
- using RCBQ with cut-decoding algorithm (with backtracking in the case of null-error),
- using Reed-Solomon code

All experiments are made for the code (72,288) with rate R = ¼. In the experiments with RCBQ we use the alphabet of nibbles Q={0, 1,…, 9, a, b, c, d, e, f} and a quasigroup given in [2]. In the standard algorithm for RCBQ we use pattern for adding redundancy: 110011001000000001100100010000000 110011001000000001100100010000000000000000, key of 10 nibbles, blocks of 4 nibbles and B_{max} = 4. For the experiments with cut-decoding algorithm we used the same quasigroup and value of B_{max}, pattern: 110011101100110011101100 110011000000 and two different keys of 5 symbols.

Fig. 2. Original image

In the experiments with both algorithms for RCBQ if *more-candidate error* appears then we randomly select a message from the set (reduced sets) in the last iteration and we take it as the decoded message. In the cases when null-error appears, i.e., $S_i = \varnothing$, we take all the elements from the set $S_{i–1}$ and we find their maximal common prefix substring. If this substring has k symbols then in order to obtain decoded message of l

symbols we take these k symbols and we add $l - k$ zero symbols.

In our experiments with Reed-Solomon codes (RSC) we used a shortened version of the RSC(63,27) ([6]). It is the code RSC(48,12) defined over the Galois field $GF(2^6)$ with primitive polynomial $p(x) = 1+X+X^6$ (and it has the same good properties as general RSC). This shortened RSC has the same length of the codewords (288bits) and the same rate (1/4) as the considered RCBQ.

Experimental results for packet-error probabilities (*PER*) and bit-error probabilities (*BER*) for different values of bit-error probability p in the binary symmetric channel are given in Table 1 and Table 2. In these tables PER_s and BER_s are the probabilities obtained with the standard algorithm, PER_c and BER_c – with the cut-decoding algorithm and PER_{rs} and BER_{rs} the probabilities obtained with Reed-Solomon code.

TABLE I: EXPERIMENTAL RESULTS FOR PACKET-ERROR PROBABILITY

p	PER_s	PER_c	PER_{rs}
0.03	0.0018	0.0032	0.0003
0.06	0.0338	0.0266	0.1157
0.09	0.1813	0.1282	0.7116
0.12	0.4719	0.3620	0.9748

TABLE II: EXPERIMENTAL RESULTS FOR BIT-ERROR PROBABILITY

p	BER_s	BER_c	BER_{rs}
0.03	0.0011	0.0010	0.0001
0.06	0.0251	0.0112	0.0241
0.09	0.1379	0.0570	0.1625
0.12	0.3715	0.1742	0.2565

Images obtained for the considered values of bit-error probabilities p in the binary-symmetric channel are presented in Figure 3 - 6. In these figures the images in a) are obtained after transmission through the channel without using error-correcting code. In b) and c) we give the images obtained using RCBQ with standard and cut-decoding algorithm, correspondingly. And in d) the images coded/decoded with Reed-Solomon code are given.

From the results given in Table 1 we can see that *PER* obtains the smallest value for Reed-Solomon code only for p = 0.03. While, for $p \geq 0.06$ the RCBQ give much better results, especially with cut-decoding algorithm (with backtracking in the case of null-error). But, from Table 2, we can see that the differences between the results for *BER* are not so significant. The reason for that lies in the construction of the RCBQ. Namely, for these codes, when a bit is incorrectly decoded, then almost all consecutive bits are incorrectly decoded. Consequently, we obtain empty decoding candidate set in some iteration and only a part of the message is decoded. As we explain above, in these cases in the place of the non-decoded part of the message we put zero symbols. In fact, these zero symbols are the horizontal lines that can be seen in the Figures 3b) – 6b) for standard RCBQ and Figures 3c) – 6c) for cut-decoding algorithm.

IV. CONCLUSION

In this paper we compare performances of RCBQ with the standard algorithm, RCBQ with the cut-decoding algorithm and Reed-Solomon codes for coding/decoding images. Analyzing results from Table 1 and Table 2, as well as results from Figure 1-6, we can conclude that Reed Solomon codes are the best only for small value of p (p <=0.03). For $p > 0.03$, random codes with the cut-decoding algorithm give the best results (the smallest values of *PER* and *BER*). Also, it can be seen from the visibility of images (Figures 3c) – 6c)). We have to notify that for standard RCBQ and Reed-Solomon codes, coding/decoding does not have sense for $p \geq 0.09$ since $BER > p$. For $p = 0.09$, RCBQ with cut-decoding algorithm give $BER < p$. The image in Figure 5c) is most visible than other images in Figure 5. For $p > 0.1$, for all considered codes we obtain worse *BER* than p.

From previous discussion we can conclude that RCBQ with cut-decoding algorithm (with backtracking) have the best performances for transmitting images through binary symmetrical channel.

REFERENCES

[1] D. Gligoroski, S. Markovski, Lj. Kocarev, "Error-correcting codes based on quasigroups," in *Proc. 16th International Conference on Computer Communications and Networks* (ICCCN 2007), 2007, pp.165-172.

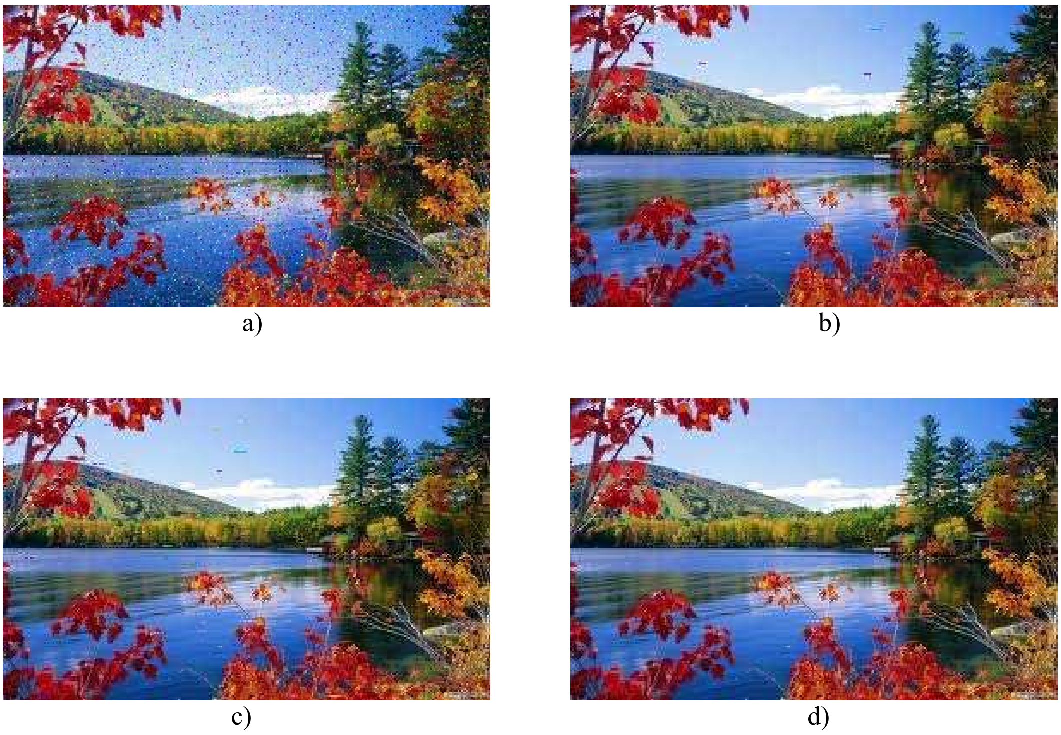
[2] A. Popovska-Mitrovikj, S. Markovski, V. Bakeva, "Performances of error-correcting codes based on quasigroups," in *ICT-Innovations 2009*, D. Davcev, J.M. Gomez, Eds., Springer, 2009, pp. 377-389.

[3] A.Popovska-Mitrovikj, V.Bakeva, S.Markovski, "On random error correcting codes based on quasigroups", in *Quasigroups and Related Systems* 19 (2011), pp. 301-316

[4] A. Popovska-Mitrovikj, S. Markovski, V. Bakeva, "Increasing the decoding speed of random codes based on quasigroups," in *Web Proc. ICT Innovations2012*, ISSN 1857-7288, Ohrid, 2012, pp. 93-102.

[5] D. Gligoroski, S. Markovski, Lj. Kocarev, "Totally asynchronous stream ciphers + Redundancy = Cryptcoding," in *Proc. of the 2007 International Conference on Security and management, SAM 2007* S. Aissi, H.R. Arabnia , Eds., CSREA Press, Las Vegas, 2007, pp. 446 – 451.

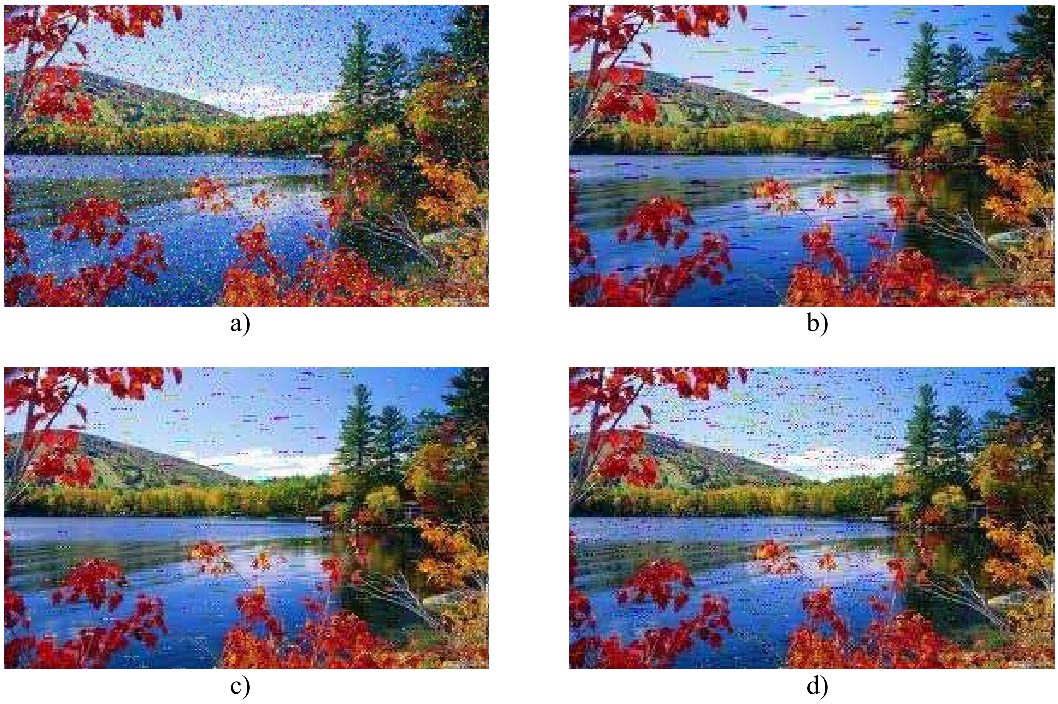
[6] J. C Moreira, P. G. Farrell, "Essentials of error-control coding", John Wiley and Sons, Ltd (2006).



a) b)

c) d)

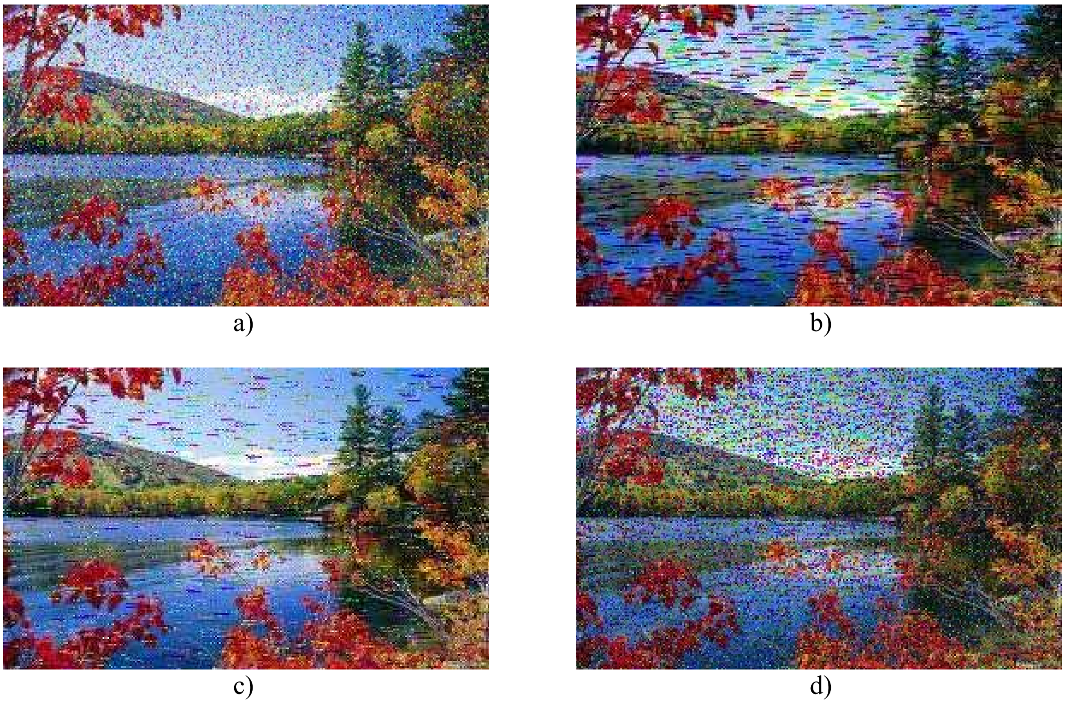
Fig. 3. Images for *p* = 0.03



a) b)

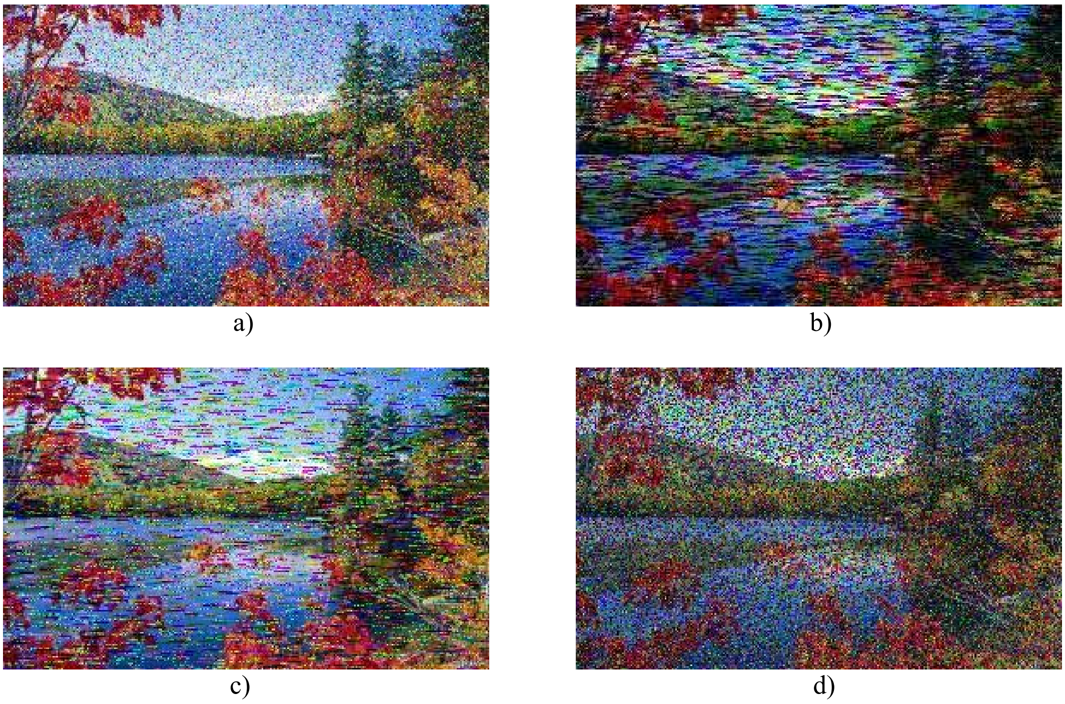
c) d)

Fig. 4. Images for *p* = 0.06



a) b) c) d)

Fig. 5. Images for p = 0.09



a) b) c) d)

Fig. 6. Images for p = 0.12

Realization Of Arithmetic Operations With a Turing Machine By Using Finite Automata

Dobre Blazhevski, Nikola Pavlov, Veno Pachovski, Adrian Bozhinovski
School Of Computer Science and Information Technology
University American College Skopje
Skopje, Republic Of Macedonia
{dobre.blazevski}{nikola.pavlov}{pachovski}{bozhinovski}@uacs.edu.mk

***Abstract*—There are several types of computational models. Finite automata are the simplest models suitable to model computers with an extremely limited amount of memory. Finite automata can be both deterministic and nondeterministic and are useful models for important kinds of hardware and software such as software for designing and checking digital circuits, lexical analyser of compilers, finding words and patterns in large bodies of text, verification of systems with a finite number of states etc. A Turing machine model is similar to finite automata, but has an unlimited memory, and can be used to compute anything that a real computer can. In this paper, first the brief theoretical background of Finite automata and the Turing machine will be exposed. Then, the focus will be switched on designing and realization of different types of arithmetic operations with a Turing machine. All of the designs will be realized with an appropriate software simulation.**

Keywords—finite automata, Turing machine, deterministic, nondeterministic, computational model, arithmetic operation

I. Introduction

As it is stated in [1], real computers are too complicated to be able to be directly modeled mathematically. That is why computational models are used instead. Finite automata and the Turing machine are one of these models.

Finite automata (FA) or Finite state automata (FSA) are the simplest computational model. They are good models for computers with extremely limited amounts of memory [1, 2]. Such computers lie at the heart of various electromagnetic devices such as the controller for an automatic door, the controller for an elevator, digital clocks etc. FA are useful models for important kinds of software such as software for designing and checking digital circuits, lexical analyser of compilers, finding words and patterns in large bodies of text, verification of systems with a finite number of states etc.

The Turing machine was first proposed by Alan Turing in 1936 [1, 3]. He proposed the Turing machine as a model of “any possible computation” [4]. The Turing machine is a more powerful model than finite automata. According to [1], it is similar to a finite automaton, but it has unlimited memory and is a much more accurate model of a general purpose computer. Therefore, the Turing machine can do everything that a real computer can do. Similarly, this model can accommodate the idea of a stored program machine like a computer [3]. The Turing machine can be used as a Language accepter that accepts a recursive or a recursive enumerable language, then as a Function transducer to compute a total recursive or a partial recursive function, and as a Decision problem solver where the machine is used as an algorithm to solve or partially solve a class of decision problems [2].

In this paper, first the brief theoretical background of Finite automata and the Turing machine will be exposed. Then a realization of arithmetic operations f(x, y) = x + y and f(x, y) = x - y with a Turing machine will be shown. All of the designs will be realized with JFLAP. JFLAP is software that allows experimenting with different types of automata, languages, parsing and L-systems [5]. The final section will contain a conclusion based on the work done and future work.

II. Finite Automata

A. State Diagram

Fig. 1 presents the state diagram of a finite automaton named M1. The circles present states. As can be seen in Fig. 1, the finite automaton in this case has three states, labeled q1, q2 and q3. Each automaton has a start state (q1) which is indicated by the arrow that points at it from the outside, and an accept state (q2). The accept state is the one with double circle [1, 6]. The arrows that lead from one state to another (or from one state back to the very same state) are called transitions [1].

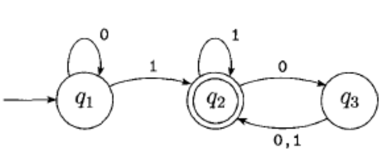


Fig. 1. State diagram of the Finite automaton M1 [1]

According to [3], a FSA can be seen as an oriented graph with labels on each arc, whereas, according to [6], a FA is a directed graph that consists of a finite set of labeled circles called states and a finite set of arrows. The automaton processes the string that appears on its input. Let’s assume the string to be 1101. The processing begins in the start state q1. The automaton reads symbols from the input string one by one and from left to right [1]. After reading each symbol, it moves from one state to another, according to the transition that has that symbol as its label [1]. In this case, the processing starts in the state q1. Then automaton reads 1 and follows the transition from q1 to q2. Then it reads the second 1 from the input string and as a result it follows the transition from q2 to q2. Then the automaton reads 0 and it follows the transition from q2 to q3. When it reads the last symbol, it follows the transition from q3 to q2. As it is stated in [1], the automaton produces an output after reading the last symbol and the output is either accept or reject. In this example it is said that the automaton is in the accept state, because after reading the last symbol it is in the accept state q2. If after reading the last symbol the automaton is in another state than the accept state, then the output is reject [2].

B. *Formal Definition*

According to [1], a formal definition of Finite automata is needed because it is precise and therefore it resolves any uncertainties about what is allowed in finite automata and because it provides good notation.

A finite automaton has a set of states [1, 4], a start state and a set of accept states. It has an input alphabet that indicates the allowed input symbols. Then, it has rules for moving from one state to another, depending on the input symbol. Therefore, according to [1], the formal definition of finite automata says that it is a 5-tuple consisting of set of states, input alphabet, rules for moving, start state, and accept state, i.e. a finite automaton is a 5-tuple (Q, ∑, δ, q0, F), where

1. Q is a finite set called the states,
2. ∑ is a finite set called the alphabet,
3. δ: Q x Z→Q is the transition function,
4. q0 ∊ Q is the start state, and
5. F ⊆ Q is the set of accept states.

The transition function δ defines the rules for moving. As an example, if the finite automaton has a state q1 with an arrow to a state q2, that is labeled with the symbol 1, which means that if the automaton is in the state q1, when it reads 1, it moves to the state q2. The transition function δ specifies exactly one next state. Having in mind the above, the formal definition of the finite automaton presented in Fig.1 will be as follows:

M1 = (Q, ∑, δ, q1, F), where

1. Q = {ql, q2, q3},
2. ∑ = {0, 1},
3. δ is described as

	0	1
q1	q1	q2
q2	q3	q2
q3	q2	q2

4. q1, is the start state, and
5. F = {q2}.

The language of the machine M is the set of all strings that the machine accepts

$$L(M) = A \tag{1}$$

In equation (1), L is the language that machine M recognizes and A is the set of strings that the machine M accepts. A finite automaton may accept several strings, but it recognizes only one language [1]. The finite automaton M1 presented in Fig.1 recognizes the set of strings that contain at least one 1 and an even number of 0s following the last 1. Therefore M1 recognizes the string 1101 as well.

Deterministic Finite Automata (DFA) has only one transition after reading the sequence of inputs from each state [3]. The finite automaton M1 described above is deterministic, because when it is in a given state and reads the input symbol, the next state is determined [1]. However, if several choices exist for the next state at any point, then automaton is called a Nondeterministic Finite Automaton (NFA). The nondeterminism allows the machine to select arbitrarily from several possible responses to a given situation [2].

C. *Turing machine*

Before the existence of the computers in 1930's , A. Turing studied an abstract machine that had all the capabilities of today's computers. The goal of Turing was to describe precisely the boundary between what a computing machine could do and what it couldn't do. His conclusions apply not only to his abstract Turing Machines (TM), but to today's real machines as well. The purpose of the theory of undecidable problems is not only to establish the existence of such problems, but to provide guidance to programmers about what they might or might not be able to accomplish through programming. In other words, it can be said that any problem that is algorithmically solvable and has an efficient algorithm for the related solution, can also be solved and demonstrated using a Turing Machine.

The formal notation that will be used for a Turing machine is similar to the one used to describe finite automata. The TM can be described by the 7-tuple [2]

M = (Q, ∑, Γ, δ, q0, B, F), the components of which have the following meaning:

Q: The finite set of states;

∑: The finite set of input symbols;

Γ: The finite set of tape symbols; ∑ is always a subset of Γ. The number of symbols in Γ ≥ ∑.

δ: The transition function. The arguments of δ(q, X) are a state q and a tape symbol X. The transition function can be defined as triple as well, i.e. δ(p, Y, D), where:

- p is the next state, in Q:
- Y is the symbol, in Γ, written in the cell being scanned, replacing whatever symbol was there.

- D is a direction, either L or R, meaning "left" or "right" respectively, informing us about the direction in which the tape head moves.

q0: The start state, which is a member of Q, where finite control is found initially.

B: The blank symbol. This symbol is a part of the Γ, but not of the ∑ subset. This is not an input symbol. The blank appears initially in all but the finite number of initial cells that hold input symbols.

F: The set of final or accepting states, which is a subset of Q.

Fig.2 presents the transition from q1 to q2 where the TM reads from the tape and writes. *L* and *R* represent the movement of the head to the left or to the right.

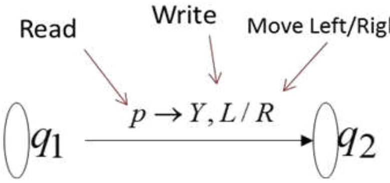


Fig. 2. Transition of TM [7]

Similar to finite automata, but with unlimited and unrestricted memory, a Turing machine is a much more accurate model of a general purpose computer. It contains unlimited tape as its unlimited memory and a tape head which can write and read to the tape using symbols in the alphabet ∑. Fig.3 presents the tape of a Turing Machine and a head for reading and writing. The head can also move around the tape. The power of a Turing machine is limited due the power of an algorithm. Initially, the tape contains only the input string and is blank everywhere else. If the machine needs to write some data, it can write them on the tape. To read that data, the machine can move the tape head over it. The TM continues to compute until it produces an output. The outputs accept and reject are obtained by entering those states respectively. If it doesn't enter either of these states, it will go on forever, never halting.

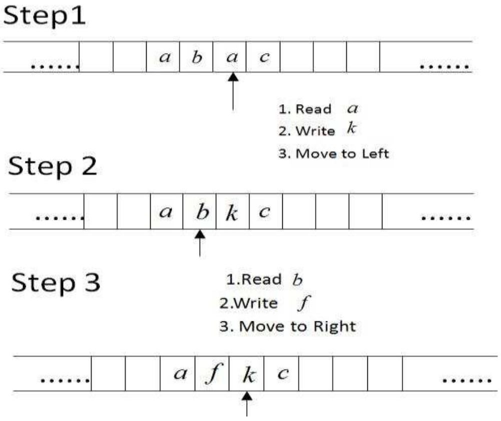


Fig. 3. TM tape and head for reading and writing [7]

The collection of strings that M accepts is the language of M, or the language recognized by M, denoted as L(M).

Accepting and rejecting configurations are the halting configurations and those do not yield any further transitions, meaning after the accepting or rejecting state is reached, the machine stops working i.e. it halts. No other transitions are possible from these configurations. What can be concluded from this, is that the input string is acceptable if and only if the machine M goes into the state q_accept, which is an accepting state. Otherwise it can be said that the input string is not acceptable if M goes to state q_reject, meaning there is no transition available for some symbol of the input string, or it can happen if M goes in an infinite loop and never ends. The machine then will stop working and go to q_reject [2].

D. Determinism

The basic model of the Turing machine is known to be deterministic. This can best be described by showing what kinds of transitions are allowed inside a Turing machine and what are not allowed. Let's assume that there is an initial state q1 and a transition leading to two other finishing states q2 and q3. Lets assume that these two transitions are as follows:

δ (q1, a) = (q2, b, R) and δ (q1, a) = (q3, d, L),

In this case the Turing machine will not know what to do, choosing only the first transition, leaving the second one like it never existed. This situation is illustrated in Fig.4 as allowed and not allowed transitions of Turing machine.

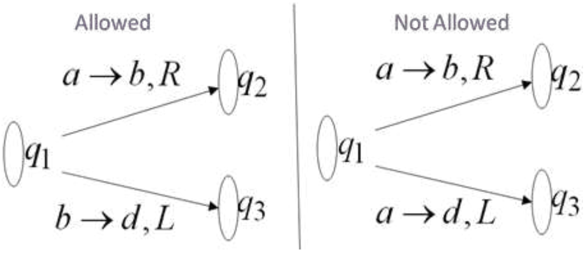


Fig. 4. Allowed and not allowed transitions in a TM [7]

E. Variants of Turimg Machine

Because the TM is a hypothetical model and not a real machine, there can be other variants of it instead of just the basic model. The original model and its reasonable variants all have the same power, meaning they recognize the same class of languages. Here some of those variants and their equivalence in power will be described. To demonstrate these variants, the type of the permitted transition function can be varied. As explained in the formal definition of the TM, the transition function forces the head to move left or right. In one variant, another property instead of just R or L can be used, called "stay put" S. This transition will allow the tape head not to move during the transitions. This doesn't give the machine any more power, because the stay put property can be demonstrated also with using one both the L and R property once more [1]. The most known variants of Turing machines are:

1) Multitape Turing machine

It has a multiple number of tapes. Each tape has its own head for read/write. Initially the input string shows on tape number one, while the other ones are empty.

The transition function can be explained with:

δ: Q x Γ > Q x Γ(k) $\{L,R,S\}^K$,

where k is the number of tapes.

Each Turing machine with more tapes can be converted to a single tape TM [1].

2) Nondeterministic Turing machine (NDTM)

It uses two kinds of search strategy [1]

- Breadth first search
- Depth first search

3) Enumerators

Turing Machines with a printer attached on them. Each input string that they work with is printed on the printer [1].

F. Unary Number System

The unary number system is a numbering system best used for solving problems using Turing machines. It contains only the symbol 1 [7]. For example, the number 5 in the decimal system will be represented as 11111 in the unary numbering system. The number of 1's in unary is equal to the size of the number n in binary. Using this system, the work of Turing machines is greatly simplified.

III. Examples

As it has been said, the TM has an ability to represent any efficient algorithm. Its power its limited due to the power of the algorithm, so if there is a known existent solution, it can surely be presented with the Turing machine. The following example will demonstrate how the Turing machine works with simple arithmetic calculations. In this example, addition and subtraction will be demonstrated.

A. Addition

Fig. 5 shows a TM that is used to obtain the sum of two numbers x and y.

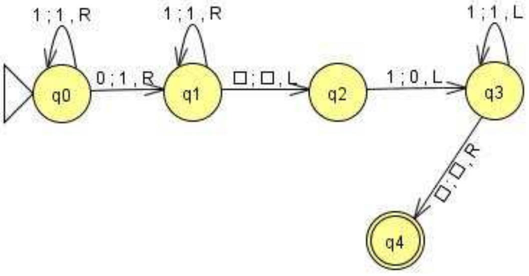


Fig. 5. State diagram of TM in JFLAP that calculates f(x, y) = x + y

The algorithm was designed in JFLAP according to its capabilities [5]. The machine has 5 states {q0…q4}. It's noticeable that each state has one or more transitions, except for the q_accept or accepting state which, in this case, is q4. Thus, the machine can take the form:

M = (Q, Σ, Γ, δ, q0, □, F)

where

Q = {q0, q1, q2, q3, q4}

Σ = {1, 0}

Γ = {1, 0, □}

δ is defined as follows:

δ	1	0	□
q0	(q0, 1, R)	(q1, 1, R)	/
q1	(q1, 1, R)	/	(q2, □, L)
q2	(q3, 0, L)	/	/
q3	(q3, 1, L)	/	(q4, □, R)
q4	/	/	/

F = {q4}.

Σ = {1, 0} means that only acceptable input for this machine will be 1 and 0. On all other symbols, the machine would crash and go to rejecting state. It is the only acceptable language. This is because for the current example the TM is using only unary numbering system, meaning only 1's. The 0 is used as a delimiter to separate the two numbers that should be summarized.

As it was described in [7], the input needs to have the form x0y (which means x + y) and the output should be xy0, which in unary numbers should give the result. δ is the transition function. It tells about each of the properties of the transition function. After the input is entered, M starts to work step by step. The input is inserted on the tape and the machine starts with the initial state q0. The whole point is that the 0, which is the delimiter, is put on the right end of the input. That way the format xy0 will be obtained.

Starting from q0 there are two transitions. If the head reaches 0 in the input, it will go to state q1, and if it reaches 1 it will stay in the same state, meanwhile going 1 step to the right on the tape. q1 is similar to q0. On each transition, the first symbol is the "if statement", the second is the replacing symbol and the third is telling the direction in which the head should move {L/R}.

B. Subtraction

In this example function f(x, y) = x – y will be demonstrated. If x > y, then the output will be (x – y) whereas if x < y, then the output will be –(x – y). If x = y then the output will be a blank tape. In order to build a TM that can calculate the function from above, the approach from [8] will be used. The input string of the TM is presented with the x and y portion. x and y are separated with 0. The approach removes one 1 from the end of the string, i.e. from the y portion. Then, it removes 1 from the beginning of the string, i.e. from the x portion. The operation continues until there are no more 1s in the y portion. In this case (x - y) will remain on the tape.

In addition to this approach, if there are no more 1s remaining in the x portion, then x < y. In this case the 0 is cleared from the tape and "–" is written in front of the remaining 1s. This way one possible solution of –(x – y) is represented, that will remain on the tape. Moreover, in the case where x = y, then there will be a blank tape left. The TM will be defined as follows:

M = (Q, Σ, Γ, δ, q0, □, F)

Q = {q0, q1, q2, q3, q4, q5, q6, q7, q8}

Σ = {1, 0, -}

Γ = {1, 0, -, □}

δ is defined as follows:

δ	0	1	–	□
q0	(q7, -, R)	(q1, 1, R)	/	/
q1	(q1, 0, R)	(q1, 1, R)	/	(q2, □, L)
q2	(q5, □, L)	(q3, □, L)	/	/
q3	(q3, 0, L)	(q3, 1, L)	/	(q4, □, R)
q4	/	(q0, □, R)	/	/
q5	/	(q5, 1, L)	/	(q6, □, R)
q6	/	/	/	/
q7	/	(q6, 1, L)	/	(q8, □, L)
q8	/	/	(q6, □, R)	/

F = {q6}.

Fig.6 presents the state diagram of the TM that calculates f(x, y) = x - y, where x > y, x < y and x = y. The TM was designed in JFLAP [5].

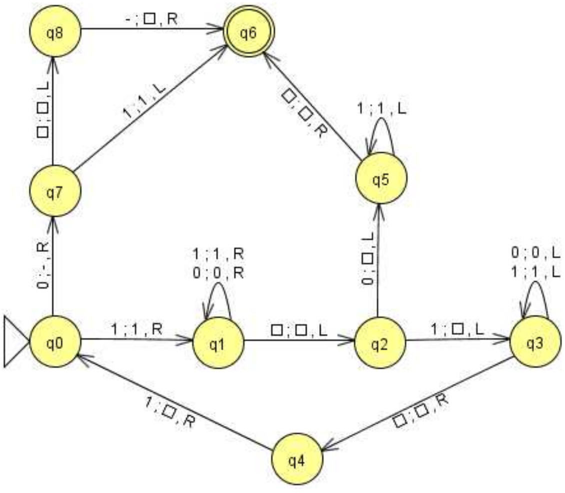


Fig. 6. State diagram of TM in JFLAB that calculates f(x, y) = x - y

IV. Conclusion

Real computers are too complicated to allow us to set up a manageable mathematical model of them directly and therefore computational models are used instead. Finite automata (FA) are the simplest computational models and they are good models for computers with extremely limited amounts of memory. On the other hand, A Turing machine is a similar but more powerful model than finite automata. It has unlimited memory and is a much more accurate model of a general purpose computer. A Turing machine can be used as a Language accepter, Function transducer and Decision problem solver. In this paper, JFLAP was used as a simulation tool in order to show how a Turing machine computes the f(x, y) = x + y and the f(x, y) = x – y.

As future work, implementation of f(x) = 2x, f(x) = 3x, f(x) = 4x, as well as more appropriate approach to f(x) = 16x is considered to be done.

Acknowledgment

In this paper JFLAP was used in order to present how a Turing machine works. We thank JFLAP.

References

[1] M.Sipser. Introduction to the Theory of Computation. Second Edition. Boston, Thomson, 2006

[2] S.Y. Yan. An Introduction to Formal Languages and Machine Computation : Principles and Practices. Fourth Edition. Singapore, World Scientific, 1998

[3] R. Roche and Y. Schables eds. Finite-State Language Processing. Cambridge, A Bradford Book The MIT Press, 1997

[4] J.E. Hopcroft, R. Motwani and J.D. Ullman. Introduction To Automata Theory, Languages, and Computation. Second Edition. Addison-Wesley, 2001

[5] JFLAP. "JFLAP", 2009, available at: http://www.jflap.org [last accessed on February 11, 2014]

[6] G. Tourlakis. Theory of Computation. NJ, Wiley, 2012

[7] College2Day. Study Matterial Examples, available at: http://college2day.com/jiit/fms/sm/CSE_IT/5th%20Sem/Theory%20of%20Computation/TOC%20MATERIAL%20GS%20AFTER%20T2/Chap3%20-%20Turing%20Machine%20Examples.ppt [last accessed on January 30, 2014]

[8] C. Colbourn. Homework Five Solution-CSE 355, 2012, available at http://www.public.asu.edu/~ccolbou/src/355hw5s12sol.pdf [last accessed on December 3, 2013]

Session 10

Doctoral research

A new framework of QoS-based web service Discovery and Binding

Festim HALILI and Merita KASA HALILI
Department of Informatics
State University of Tetova
Tetovo, Macedonia
festim.halili@unite.edu.mk

Abstract—**The deployment of similar or same web services into dedicated repositories (or registries) for the end-users to discover and select them has significantly increased lately, which has led to a new problem of choosing a suitable web service for the requester based on their needs and expectations. In view of this challenge, a QoS-based method for web service ranking is proposed in this paper. To achieve desired selection of a quality service, it is necessary to produce a framework that enables evaluation of a web service quality based on factors in terms of attributes, in addition to their level of importance based on the opinion of skilled engineers of Service Oriented Computing. The primary goals are matchmaking, qualifying, categorizing and ranking services based on proposed factors, taking both functional and non-functional properties of the services to be evaluated. Based on the proposed factors, the service providers can improve qualities of their web services, whilst the requester will have a proper list of ranking.**

Keywords—web services; QoS; service discovery; ranking; selection; quality factors; SOC

I. INTRODUCTION

The traditional approaches to software development where interactions are based on the exchange of the products with specific clientship are known as object-oriented programming, but they are not designed to face the challenges of open environments. SOC (Service Oriented Computing) [1] is a new paradigm and provides a way to create a new architecture that is determinant toward autonomy, versatility and heterogeneity. SOC uses services to sustain the development of low-cost, loosely-coupled, evolvable and massively distributed applications. Services are autonomous and self-describing entity and they can be published and discovered by the service requestor. The deployment of web services into traditional registries such as UDDI [4, 5] is massively executed by service providers. According to Zheng [6] by 2010 there have been existing around 28.500 public web services on the internet, and it is clearly that the number has increased by now.

Web services are versatile, meaning that they can be used in different styles like: RPC, REST and SOA [7, 8] and the benefits of using them by business or non-business companies is explained in [9,10]. The architecture of web services relies on different standards or protocols, the one for service description based on functional properties is WSDL [11], but the lack of this standard is that it doesn't evaluate or describe non-functional properties of web services, by means of QoS attributes, as it can be seen in the figure 1, where the Abstract part describes the sent and received messages and the operation which associates a message exchange pattern with one or more messages, whilst the Concrete part specifies the transport and wire format details for one or more interfaces and also the port (an endpoint) associates a network address with a binding.

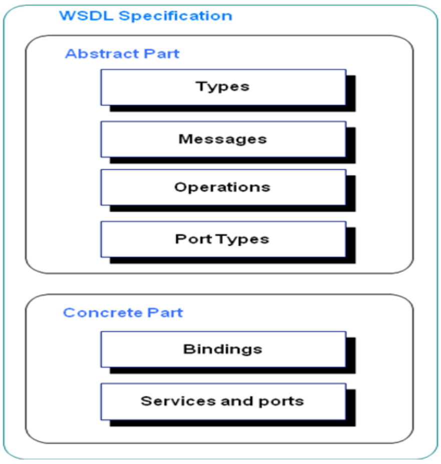


Fig. 1. The WSDL Document Structure

To define the quality of web services, it is a difficult target; because the quality may depend from task-related factors that would affect the end users requirements.

A true concern to the service selection is the preliminary way to rank the matching services based on some criteria: a) semantic-based measure [13] or non-functional properties (QoS or eminence) [14]. There are proposed technical and methodical aspects for understanding the engineering foundations of SOC that concern the way applications can be developed to provide solutions to their business, without the regard of the programming languages the services are coded. SCA (Service Component Architecture) is proposed by the Open Service Oriented Architecture collaboration which addresses the aim of creating the model for service components. However, it points out a low-level design [2] in terms of assembly model and binding mechanism. Another modeling language proposed by SENSORIA is SRML (The SENSORIA reference modeling language) [3, 12] which conveys both the static and dynamic aspect, where the static aspect includes the design-time description of complex services in terms of composition (orchestration) of simpler services which are indicated over state transition systems and temporal logic, whilst the dynamic aspects are formalized over a specified mathematical model. In this paper we will aim to extend the abstract model of service discovery and binding proposed in SENSORIA by including specific QoS factors in the part of Configuration Management of the proposed architecture, which will assist in the ranking process of web services. According to [2] the ontology unit in their proposed framework is an area that is still lacking standards, even though there has been perceptibly progress in the development of Semantic Web Techniques.

The current proposed selection systems concerning the requester (user) design have few issues left aside. Firstly, current systems are presuming that users have the abilities of formulating queries that can reflect their QoS requirements, because users might have no knowledge about what the realistic QoS values indicate. It is very important if the selection systems can provide to users the ability to choose the right QoS values for their requirements.

So, our aim in this paper will be to provide an advanced system with QoS factors as ranking criteria, where users could go through all available services in the given registries (UDDI) to gain particular ideas on their QoS value ranges.

II. Related Work

In order to discover Web services, it is necessary to obtain a collection of available Web services. The QoS has been previously exploited to support the service selection. In the beginning stage of Web services research, UDDI (Universal Discovery Description and Integration) was proposed as a repository or registry for publishing and invoking services, but it has not prevailed in the area of publicly available Web services [15] and by time passing other systems were proposed.

A prerequisite to selection is the process of ranking the matching services based on some criteria: a) semantic-based measure or b) non-functional properties (QoS or eminence), and some ideas in different papers are proposed based on these methods.

A. Semantic-based measure

In [2] it is proposed a unification process of discovery, ranking and selection based on the formal operational semantics. In addition, for modeling this process, there are used unifying wires for a given business configuration system. The semantics is based on a graph-based representation and configuration of the GC (global computers) specified by business activities. The three steps: discovery, ranking and selection are based in compliance with required business and interaction protocols and a tentative of optimization of QoS constraints, but there are not specified exact QoS factors to determine the ranking of available web services.

In [16], it is proposed to add context to knowledge of service provider and requester to service description and incorporate a planning method with intention to achieve the understanding for the inner ontology concepts between parties in service binding.

Segev and Toch in [17] indicate a two-step context-based semantic approach to the problem of binding and ranking Web services for possible composition of services (orchestration or choreography [9]), and it is also given an empirical analysis and comparison of the different methods for classification (WSDL context, WSDL TF/IDF and Description context).

In addition, the semantic-based measure advantage is that many of the tasks involved in using Web services can be (semi) automated, including discovery, selection, composition, mediation, monitoring and ranking. However, many systems have been developed without considering the ability of the Web for integrating services and sharing resources, and thus the industrial rise of Semantic Web Services technologies has been slower than predicted.

B. Non-functional properties

The process of selecting the suitable Web service to a client requirement is an important task as many Web services are able to meet functional requests of user, however, non-functional attributes is an onerous matter also. Various non-functional properties based on QoS models have been proposed in the literature in which QoS parameters are evaluated by a service broker or matchmaker. One such model is evaluated in [18], where the QoS consultant acts as a broker or agent between client and service provider. Once, a search is performed for query, the mediator offers to the requester a list of candidate services that are matched with the given request.

In [5] it is proposed the real-time assurance of service responsiveness incorporated in UDDI registries, but it is not evaluated the process of ranking services according to specific requirements of the service searcher or client, that would ease the process of selection, which is the intention of this paper.

In [19, 20] are proposed QoS based architectures for ranking and selection of web services, by describing the capabilities of a Web service to fulfill the requirements of consumers in specific domain.

Jaeger and Ladner in [21] proposed improving method for QoS of Web Service compositions based on redundant services. They discussed how already identified candidates which a selection process originally has divided out, can lead to improvement of composition with particular QoS categories. They have proposed four QoS categories: execution time, cost, reputation and availability, but they didn’t evaluate other categories, because service requesters often cross the lines of marginalized categories.

An interesting approach to non-functional properties is proposed by Ahmadi and Binder [22], by providing a flexible matchmaker for service discovery, selection and ranking, including both, functional and non-functional properties into consideration. In the proposed framework, the matchmaker provides an expressive language for the clients to define service requests, by indicating the involved registries or service repositories, non-functional parameters and an utility function for ranking Web services. Here, the UDDI are separated from the third party repositories and the matchmaker plays a crucial role enabling the service requestors to search inside them, while supporting different service description languages and emerging languages. However, perception of the structural properties of frameworks often assists in gaining better insights and develops better algorithms. Based on this, in our paper, relied on large-scale services discovery, we propose an uninterrupted method to rank web services.

III. Proposed Service Ranking and Selection Framework

A. *Problem Complexity*

The traditional service discovery model is known for having difficulties in the precise classification and ranking of the list of similar services published by different providers, because it is limited to only three roles: service consumer, service provider and UDDI registry as shown in the figure 2 below:

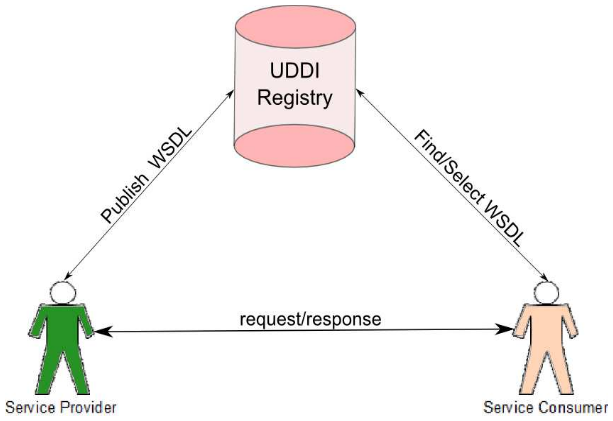


Fig. 2. Tradtional discovery of services

As we can see from the above picture the procedure of discovery and selection of services is simple and easy to understand, but when a list of services meeting consumer’s functional requirements have been discovered, it is difficult for the service consumer to choose which one should be invoked among all these services with similar descriptions and capabilities.

B. *Proposed Framework*

Our proposed framework deals with a set of nonfunctional properties indicated by QoS and it flanks performance attributes. The Service Consumer is always concerned about the performance of the service to be invoked; therefore we incorporate QoS in our framework to rate functionally similar web services. Our idea is to facilitate service providers to publish information with QoS, in order to do that; we need to model service descriptions. Furthermore, it is necessary to provide a method for consumers to submit service requirements according to their needs. A powerful web service modeling language has been given in [3] called SRML (Sensoria Reference Modeling Language). Nevertheless, it does not support QoS registry and attributes, therefore we propose a service reference modeling language including QoS named SRML-Q. We extend the configuration management of the overall ‘engineering’ architecture and processes proposed in [2,3] by including several components to support service classification, ranking, quality update and binding or selection. The main framework for web service selection and ranking with QoS attributes is shown below in the figure 3.

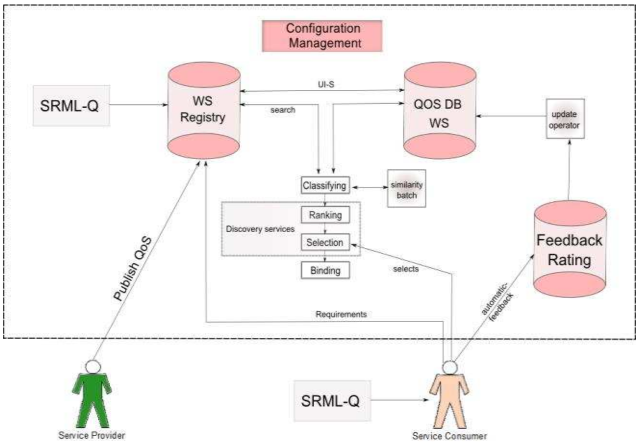


Fig. 3. Proposed QoS approach ranking and selection framework

We have defined a web service in the WS-Registry with six elements, one added to the definition in [23]:

WS={WS-Name, UI-S, WS-Desc, Qf, FSet, In}

According to the above tulpe we denote:

- WS-Name as the name of the specific web service,
- UI-S as the unique identifier of the web service, with intention to discern them from each other.
- WS-Desc service role description
- QF is the published QoS factor description that is specified as QF = QS U QD
- FSet is the web function set pointed as FSet = {fset1, fset2, …, fsetn}, where each fseti=(1<=i<=n) can be triggered for a certain operation task.
- In is the invocation factor, denoting the number of times a web service is requested and binded to a specific consumer.

Similarly we define the service request, by just removing the UI-S, which is not necessary in this case.

This system assists the Service Consumer to search web services based on the input/output operations which incorporate with the QoS requirements. The similarity batch assists in classifying similar web services and prepares them to send to the ranker, who proceeds in ranking web services based on the QoS preferences set by the client and highest ranked service will be provided. There are occassions when many similar web services are readily available to meet consumer request. In case the list of the provided services by the QoS database is long, then the service selection method is activated that take inputs as matches services as per client preferences, afterwards depending on the approach of the constraint’s and client’s requests, it runs the proper web service selection algorithm and provides the results to the client.

The client has the right to rate the consumed service by providing information regarding the working experience of the service, then the UpdateOperator refreshes quality criteria value in the QoS database regarding the collected feedback information in quality rating database.

IV. Categorizing Quality Factors

To measure quality of web services, we would need to define and categorize the quality factors. In [24], the quality factors are formulated based on three objectives: Usability, Conceptual Reliability and Representation Reliability. In the appendix of the same paper, there are given tables with identified service quality factors and quality sub-factors and the number is enormous. For our framework, we propose the quality factors that are often indicated as most important to the service consumers:

Execution Time - The execution time defines the time needed to execute the service. In this paper, we presume that the values of individual services result in the overall execution time of the composition. For our calculation we consider the worst/maximal execution time as relevant.

Cost - The cost represents the amount of resources needed to use a service. For the calculation of the composition we presume that the resources are spent once a service is invoked.

Reliability – Ratio of error messages to total messages.

Reputation - The concept of reputation represents a ranking given by users of the service, like the auction platform ebay allows clients to rank the behavior of other clients [11]. Since the reputation of an individual service can be considered as the average ranking of individual users, we regard the average of the individual values as the aggregated reputation of a composition.

Availability - The availability denotes the probability that the invocation of the service performs successfully and delivers a result within the promised QoS of other categories.

V. Conclusion and Future Work

The composition of web services is faced with the large number of growing Web services into registries, there have been given a lot of efforts to deal with this problem, but still needs more research to be expressed. The process of ranking discovered Web services on the discovery service side, makes the composer able of regulating time and quality of the generated QoS based web services. In our proposed framework, we extend the traditional way of searching and selecting web services by adding QoS database incorporated with quality factors, and also proposed a SRML-Q service reference modeling language to meet the constraints between the service consumer and service provider, which would help a lot in ranking web services.

For the future work, we are going to propose several algorithms for service selection and service selection with QoS, by describing the differences between them, and also provide an algorithm to calculate and rank each service's QoS value. Furthermore, we will provide a simulation experiment of a SET of services QoS information.

References

[1] M. Huhns and M.P. Singh, “Service Oriented Computing: Key Concepts and Principles”, in IEEE Internet Computing VOL. 9, Issue. 1, February 2005.

[2] J. Luiz Fiadeiro, A. Lopes and L. Bocchi, “An Abstract Model of Service Discovery and Binding”, in Springer Verlag Journal of Formal Aspects of Computing Volume 23 Issue 4, July 2011, London Uk.

[3] J. FIadeiro, A. Lopes, L. Bocchi and J. Abreu, “The Sensoria Reference Modelling Language”, In Rigorous Software Engineering for Service-Oriented Systems, LNCS, Springer.

[4] S. Overhage, P. Thomas, “WS-Specification: Specifying Web Services Using UDDI Improvements”, Lecture Notes in Computer Science, Vol. 2593,pp. 100–119, Web and Database-Related Workshops on Web, Web-Services, and Database Systems, 2002.

[5] M.B. Blake, A. L. Sliva and M. Muehlen, “Binding Now or Binding Later: The Performance of UDDI Registries”, in pp.171c Proc. of IEEE 40th Hawaii International Conference on System Sciences, January 2007, Waikoloa, HI.

[6] Z. Zheng, Y. Zhang, and M. R. Lyu, “Distributed QoS Evaluation for

[7] Real-World Web Services,” in Proceedings of the 2010 IEEE

[8] International Conference on Web Services (ICWS ‘10), Washington,

[9] D.C., USA, 2010, pp. 83-90.

[10] F. Halili, A. Dika. Choreography of Web Services and Estimation of execution Plan, In Proc. Book. In Proc. Book of IEEE International Conference on Information Technology and e-Services (ICITeS'2012), March 2012, Sousse Tunisia.

[11] A. Dika and F. Halili. “Integrated Orchestration of Web Services and the Impact of the Query Optimization”. In Proc. Book of IEEE 8th International Conference on Computing Technology and Information Management (NCM & ICNIT), vol.2 :ICNIT Track, pp. 702-708, April 2012, Seoul, Korea.

[12] F.Halili, E.Rufati and I.Ninka “Service Composition Styles – Analysis and Comparison Methods”, in Proc. Book of IEEE CICSyN2013 5th International conference on Computational Intelligence, Communication Systems and Networks, pp.278-284, 5-7 Jun 2013, Madrid, Spain.

[13] F. Halili, A. Dika and M. Kon-Popovska. “Towards the Composition of Web Services and the Role of the Query Optimization”. In International Journal of Web Applications (IJWA), vol.4, no.2, pp.57-68, June, 2012.

[14] R. Chinnici, J.-J. Moreau, A. Ryman, and S. Weerawarana. Web services description language (WSDL) version 2.0, 2007.

[15] Fiadeiro, J.L., Lopes, A., Bocchi, L.: Algebraic semantics of service component modules. In: Fiadeiro, J.L., Schobbens, P.-Y. (eds.) WADT 2006. LNCS, vol. 4409, pp. 37–55. Springer, Heidelberg (2007)

[16] A. Yousefipour, A.G. Neiat, M. Mohsenzadeh and M.S. Hemayati, “An ontology based approach for ranking suggested semantic web services”, In Proc. Book pp. 17-22, of IEEE 6th International conference of Advanced Information Management and Service (IMS), Dec 2010, Seoul, Republic of Korea.

[17] R.J.R. Raj and T. Sasipraba, “Web service recommandation framework using QOS based discovery and ranking process”, In Proc. Book. of 3rd International Conference on Advanced Computing (ICoAC) pp. 371-377, Dec. 2011, Chennai, India.

[18] N. Steinmetz, H. Lausen, M. Brunner, “Web Service Search on Large Scale,” In Proc of ICSOC, 2009.

[19] Zh. Sheping, L. Zengzhi and W. Juanli, “A Concept Planning Algorithm for Pragmatics Web Service Discovery”, In Proc. Book of IEEE International Conference on Communication s and Mobile Computing, 2009.

[20] A. Segev and E. Toch, “Context-Based Matching and Ranking of Web Services for Composition”, in IEEE Transactions of Services Computing, VOL.2, NO.3, July-September 2009.

[21] D. Sachan, S. Kumar Dixit and S. Kumar, “A System for Web Service Selection Based on QoS”, pp. 139-144 In Proc. Book of IEEE International Conference on Information Systems and Computer Networks, 2013.

[22] T. Rajendran and Dr. P. Balasubramanie, “An Efficient Architecture for Agent based Dynamic Web Services Discovery with QoS,” Journal of Theoretical and Applied Information Technology (JATIT), Pakistan, Vol. 15, No. 2, May 2010.

[23] T. Rajendran and Dr. P. Balasubramanie, “An Efficient WS-QoS Broker based Architecture for Web Service Selection,” International Journal of Computer Applications (IJCA), Foundation of Computer Science, USA, Vol. 1, No. 9, 2010. Doi: 10.5120/194-333.

[24] M. C. Jaeger and H. Ladner, “Improving the QoS of WS Compositions based on Redundant Services”, In the Proc. Book of the IEEE International Conference on Next Generation Web Services Practices, 2005.

[25] N. Ahmadi and W. Binder, “Flexible Matching and Ranking of Web Service Advertisements”, in Proc. Book of ACM MW4SOC’07, November 26, 2007, Newport Beach , CA, USA.

[26] G. Zou, Y. Xiang, Y. Gan, D. Wang and Z. Liu, “An agent-based web service selection and ranking framework with QoS”, in Proc Book of 2nd IEEE International Conference on Computer Science and Information Technology ICCSIT, pp.37-42 , 11 Aug 2009, Beijing, China.

[27] F. Al Zaghoul, A. Al Nsour and O. Rababah, “Ranking Quality Factors for Measuring Web Service Quality”, in Proc. Book of ACM 1st International Conference on Intelligent Semantic Web Services and Applications, june 2010, Amman, Jordan.

Average Vibrational Potentials of Oscillators in Condensed-matter Environments using Hadoop

Bojana Koteska, Anastas Mishev
Faculty of Computer
Science and Engineering,
Ss. Cyril and Methodius University,
Rugjer Boshkovikj 16, P.O. Box 393,
1000 Skopje, Republic of Macedonia
Email: {bojana.koteska, anastas.mishev}@finki.ukim.mk

Ljupčo Pejov
Institute of Chemistry,
Faculty of Natural Sciences
and Mathematics,
Ss. Cyril and Methodius University,
Arhimedova 5, P.O. Box 162,
1000 Skopje, Republic of Macedonia
Email: ljupcop@iunona.pmf.ukim.edu.mk

Abstract—**In physical sciences, when condensed matter systems (e.g. solids or liquids) are modeled with an explicit inclusion of dynamical effects, often the following computational problem arises. A given property of an embedded atomic/molecular system within condensed phase should be computed either at different possible structural arrangements and further average over configurations, or alternatively, it is possible to generate an averaged configuration of the dynamical surrounding that the system experiences and further compute the property of interest at that configuration. The problem of solving the average vibrational potentials of large number of oscillators in various condensed-matter environments (sampled from a statistical physics simulation) can be placed in the category of problems with large data sets. In this paper, a distributed and parallel processing of the large data sets needed for the generation of the averaged vibrational potential is efficiently performed by using the MapReduce programming model and Hadoop software library. Some of the reasons for choosing the Hadoop software library are: It is able to work on data pieces in parallel; The computing solutions enabled by Hadoop are scalable and flexible; The distributed file system enables rapid data transfer among nodes; Hadoop is fault-tolerant which means that if a node fails the job is redirected to another node. The main goal of this paper is to perform an efficient processing of the large data sets used in the scientific applications.**

Index Terms—**Hadoop, Average vibrational potentials, Anharmonic oscillator, Condensed-matter environments, Schrödinger equation**

I. Introduction

Theoretical models in physical sciences are often used to understand the experimentally observed behavior of certain physical systems or to predict their behavior under specific circumstances which are relevant to the actual or potential technological applications of the systems in question. Besides getting a more enlightening view of the systems behavior, theoretical models may be quite useful in discriminating among various factors leading to observation of certain physical phenomena or in quantifying the contribution of various factors to a certain physical observable. Most of the experimental data are, however, collected at finite temperatures, usually quite above absolute zero.

A reliable theoretical model aiming to provide a realistic description of the system in question therefore has to account for the dynamical effects on a certain time-scale. Most of the models based on quantum mechanical description of many-particle physical systems are based on explorations of the potential energy hypersurfaces (or certain cuts through these surfaces), which means that they do not conform to the previously mentioned criterion. To explicitly include the dynamical behavior of the studied quantum system, one has to treat it within the framework of quantum dynamics. However, a fully exact quantum dynamical treatment of multi-particle systems is prohibitively computationally expensive. At the same time, luckily, such full quantum dynamical treatment is mandatory only in certain specific cases, usually when the focus of the study is put on light particles (such as e.g. hydrogen atoms).

An acceptable alternative which has been exploited to some extent in the literature is to first carry out a classical dynamics (or statistical physics, such as e.g. Monte Carlo) simulation of the time-evolution (or evolution in imaginary time) of the system in question, then to pick up a reasonably small number of configurations (snapshots from the classical simulation) and perform rigorous quantum mechanical simulations only on these configurations. Though the previously mentioned dynamical simulations are classical in a rigorous sense, note that the interaction potentials used throughout the simulations may be even derived from high-level quantum mechanical calculations.

II. Related work

There are several papers in which MapReduce paradigm has been used for solving problems in the scientific domain. In [1], the authors applied MapReduce model to perform High Energy Physics data analyses and Kmeans clustering. They also made a streaming-based MapReduce implementation and compared its performance with Hadoop. Their conclusion is that most of the scientific analyses that has some form of the SMPD algorithm can benefit from the MapReduce model and can achieve scalability and speedup.

In [2], the authors present the MapReduce implementation in Google inc. The implementation is highly scalable and it processes terabytes of data on thousands of machines. Also, upwards of one thousand MapReduce jobs are executed on

Google's clusters every day. MapReduce model is used for sorting, data mining, machine learning, generation of the data of the web server, etc.

In his thesis [3], the author propose a novel solution for molecular dynamics simulation based on Hadoop MapReduce. The solution can predict the execution time of a given size molecular dynamics simulation system. He also presents the performance and energy consumption improvement of the solution which is implemented in a hybrid MapReduce environment.

Bunch et al. [4] explore which scientific computing problems can be solved by using MapReduce and which can not. They implement different non-trivial algorithms with MapReduce and measure their performance. The authors found out that the MapReduce framework is not suitable for iterative algorithms where each iteration runs a number of MapReduce jobs.

In their paper [5], the authors propose architecture for a configuration implemented in a scientific private cloud prototype and they use Hadoop to achieve scalability and fault tolerance. The experiments showed the effectiveness of the proposed model. In [6], the authors describe the development of the Hadoop-based cloud scientific computing application that processes sequences of microscope images of live cells.

A Hadoop plugin that allows scientists to specify logical queries over array-based data models is presented in [7]. It executes queries as MapReduce programs defined over the logical data model. The goal of this paper is to reduce total data transfers, remote reads and unnecessary reads.

III. Average Vibrational Potentials of Oscillators in Condensed-matter Environments

In physical sciences, when condensed matter systems (e.g. solids or liquids) are modeled with an explicit inclusion of dynamical effects, often the following computational problem arises. A given property of an embedded atomic/molecular system within condensed phase should be computed either at different possible structural arrangements and further average over configurations, or alternatively, it is possible to generate an averaged configuration of the dynamical surrounding that the system experiences and further compute the property of interest at that configuration.

For example, if one is interested in an anharmonic oscillator embedded in a solid or liquid, the vibrational potential of the form:

$$V(r) = V_0 + 1/2 k_2 r^2 + k_3 r^3 + k_4 r^4 + k_5 r^5 \qquad (1)$$

may be computed at n configurations and then the vibrational Schrödinger equation solved for each particular $V_i(r)$. In previous equation, r is an appropriately chosen vibrational coordinate, k_2 is the harmonic force constant, while k_3, k_4 and k_5 are cubic, quartic and quintic anharmonic force constants respectively [8][9]. In these papers this approach and generated the vibrational density of states for a number of X-H oscillators embedded in a variety of liquid environments have been exploited. Though such approach is computationally feasible, in some cases, especially if one is interested only in the average frequency (or frequency shift), it would be desirable to avoid explicit computation of vibrational frequencies by solving the vibrational Schrödinger equation for all $V_i(r)$. Instead, one could use a single computation of this type, for an averaged configuration or averaged potential within the condensed phase medium. In the present study, we further elaborate the previous two ideas, by considering the averaged vibrational potential instead.

Alternatively to the averaged configuration or averaged environmental potential approaches, one can generate an averaged vibrational potential of the form:

$$<V(r)>=<V_0> +1/2 <k_2> r^2 + <k_3> r^3 + \\ <k_4> r^4 + <k_5> r^5 \qquad (2)$$

(where $<>$ denotes ensemble averaging or averaging over time configurations) and subsequently solve the vibrational Schrödinger equation for such averaged potential energy function. To illustrate the concept and consider a particular physical system, we consider the fluoroform-dimethylether dimer embedded in liquid krypton, which has been a subject of attention in our recent paper. The main interest for this system, which has previously been studied by cryospectroscopic techniques, is driven by the peculiar behavior of the C-H vibrational mode of the fluoroform moiety upon complexation with dimethylether, that exhibits C-H stretching frequency blue shift (instead of the expected red shift by ''chemical intuition''). The details concerning the mechanism behind the blue shift and many other aspects in this context have been discussed in details in our previous work.

In the present paper, we focus on the development of method, based on the map-reduce computational approach, to extract the ''solvent-averaged'' X-H stretching vibrational potential. We have therefore computed the vibrational potential energy functions for at least 50 C-H stretching oscillators of the CF_3H moiety within the CF_3H - $(CH_3)_2O$ dimer at B3LYP, HF and MP2 levels of theory. The 6-31++G(d,p) basis set has been used for orbital expansion in all calculations. The positions of the C and H atoms in the course of ''excitation'' of the C-H stretching vibration have been generated by fixing the center-of-mass of the C-H bond fixed, as explained in details elsewhere. 20-point grids were used to scan the C-H stretching potential energy function, spanning a suitable range of C-H distances, so that the potential is sampled in the areas in which the wavefunctions corresponding to the ground and the first excited vibrational states are already decayed to zero. The data generated in such way were further interpolated by a fifth-order polynomial in the C-H distance r (Eq. 1) . The functions of the form Eq. 1 were further cut after the fourth-order term, transformed into Simons-Parr-Finlan type coordinates $\rho = r - r_e/r$ (where r_e is the equilibrium value of the C-H distance), and the vibrational Schrödinger equation was solved by the variational method (the linear variant). For that purpose, harmonic oscillator eigenfunctions were used as an orthonormal basis set. To generate the averaged potential of the form Eq.2 by the map-reduce technique, we have averaged the values of molecular potential energies (in Born-Oppenheimer sense) at each r(C-H) value. The resulting average vibrational potential energy function of the form Eq.2 was further also cut after fourth order, transformed into SPF-type coordinates, and subsequently the vibrational Schrödinger equation was solved in a variational manner. The map-reduce approach, as implemented in Hadoop, was used as explained below.

IV. The MapReduce Model

MapReduce is a programming model for processing large data sets in parallel [10]. The partitioning of the input data and the scheduling of the program's execution across multiple machines are responsibilities of the run-time system. The user should specify a map function (mapper) which processes key-value pairs and a reduce function (reducer) which merges all the intermediate values associated with the same intermediate key [2].

The MapReduce model can be divided into rounds, each containing three phases: *Map*, *Shuffle and Sort* and *Reduce*. The *Map* phase maps each single pair of (key, value) to the machines in the run-time system as a new multiset of (key,value) pairs where the value in each new pair is a substring of the original value. The *Shuffle* phase is responsible for sorting and transferring the map outputs to the reducers. The *Reduce* phase computes some function on the data on each machine [11].

A MapReduce program consists of finite sequence of rounds specified as 2-tuples (tuples of two elements), each tuple containing a map and a reduce function. Formally, this can be written as: $((M_1, R_1), (M_2, R_2), ..., (M_n, R_n))$ where M_i is a mapper, R_i is a reducer, i is an integer number and $1 \leq i \leq n$. A 2-tuple is defined as (M_i, R_i). Let the program input that is a multiset of (key;value) pairs be denoted by U_0 and the output that is a multiset of (key;value) pairs of the i-th round by U_i.

The program executes for $r = 1, ...n$. For each r, the *Map*, *Shuffle* and *Reduce* phase are performed. The *Map* phase feeds each (key;value) pair $(k; v)$ in U_{r-1} to the mapper M_r and runs it. The output of the mapper M_r will be a sequence of (key;value) pairs $(k_1; v_1)$, $(k_2; v_2)$,... and it can be defined as: $U'_r = \cup_{(k;v) \in U_{r-1}} M_r((k; v))$. The *Shuffle* phase constructs $V_{k,r}$ (values such that $(k; v_i) \in U'_r$) from U'_r for each k. The *Reduce* phase feeds the k and some arbitrary permutation of $V_{k;r}$ to the separate instance of the reducer R_r and runs it for each k. The output of the reducer is a sequence of 2-tuples $(k; v'_1)$, $(k; v'_2)$,... and U_r that is a multiset of (key;value) pairs produced by the reducer R_r is defined as $U_r = \cup_k R_r((k; V_{k,r}))$ [12] [13].

Programs that use the MapReduce model implement the Mapper and Reducer interfaces to provide the map and reduce functions. The map and reduce methods can be represented as shown below. The values with the same key are reduced together [14].

method Map(key k, value v) $\rightarrow$ EMIT(key k', value v')

method Reduce(key k, value v) $\rightarrow$ EMIT (key k', value $[v', v_2, v_3...]$

The MapReduce model automatically supports parallel programming and it shields the programmer from writing code about data distribution, scheduling and fault tolerance. The programmer should only specify the map and reduce functions. This also can be considered as a disadvantage of the model since the programmer cannot affect the efficiency of the parallelism. Thus, it is not always clear which kind of problems are suitable to be solved using the MapReduce model and which not [13]. The scientific data volumes and clustering algorithms used in chemistry, biology, physics are computing intensive operations and the use of parallelization techniques is key in order to achieve efficient data analyzes. MapReduce model is suitable when processing of the data should be split into smaller independent computations and the intermediate results should be merged after some post-processing in order to get the final result. It provides simplicity, robustness and has less synchronization constraints which supersede the additional overheads [1].

Hadoop (Apache Hadoop) is an open source software data-processing library which allows distributed and parallel processing of large data sets. The Hadoop project includes four different modules: Hadoop Common (utilities that support the other Hadoop modules), Hadoop Distributed File System (distributed file system that provides high-throughput access to data), Hadoop YARN (framework for job scheduling and cluster resource management) and Hadoop MapReduce (A YARN-based system for parallel processing of large data sets) and it is used by many companies including Facebook, Cloudera, Amazon, Microsoft, Yahoo, etc. The data processing in Hadoop can be implemented in MapReduce directly or by using high-level languages and translating into Map-reduce jobs later [15][16]. Hadoop can be also used for building data warehousing solutions. An example is Hive which supports queries expressed in a HiveQL (SQL-like declarative language). The queries are compiled into map-reduce jobs and executed on Hadoop [17].

V. Computing Averaged Vibrational Potentials Energies by using Hadoop

The purpose of the our algorithm is to compute the average vibrational potential energies for at least 50 C-H stretching oscillators of the CF_3H moiety within the CF_3H - $(CH_3)_2O$ dimer at B3LYP, HF and MP2 levels of theory. Calculating the average values, in our case the average vibrational potential energies, is a typical map-reduce problem. Each document that describes the same C-H stretching oscillator in a different environment has two columns, one containing the distances r(C-H) and the other containing values of molecular potential energies (U). The pseudo code of the Map and Reduce methods used in our algorithm is given below.

```
Method Map(String r, String U):
// r: input key r(C-H)
// U: input_value
```

```
for each r in all documents:
EmitIntermidiate(r,ParseDouble(U));

Method Reduce(String r, Iterator interm_vals):
// r: key, same as input_key
// interm_vals: intermediate values-
// list of all U-s group by r
double sum=0,result=0;
for each v in interm_vals:
sum += v;
result=sum/length(interm_vals);
Emit(AsString(result));
```

The algorithm is implemented in Java and it was performed three times, once for each level of theory (B3LYP, HF and MP2).

The main results from the present study are summarized in Fig. 1 a-c. In Fig. 1, the vibrational density-of-states (DOS) histograms generated from the computed vibrational frequencies of the $|0 > \rightarrow |1 >$ C-H stretching vibrational transition are presented, together with the delta-like function (with dashed lines) representing the frequency of the $|0 > \rightarrow |1 >$ transition obtained for an averaged C-H stretching potential. In the same figure, also the numerical values of the corresponding frequencies are given. As can be seen, if one is interested solely in the average vibrational frequencies, and not in the corresponding distributions, our "averaged vibrational potential" approach gives excellent results. The matching between average vibrational frequencies computed from the DOS distributions, and the single vibrational frequency values computed from only a single averaged vibrational potentials is excellent, regardless on the level of theory. No biasing effects are present.

VI. Conclusion

In this paper we have presented the benefits of using the MapReduce model and Hadoop framework in scientific domains. The average vibrational potentials of oscillators in condensed-matter environments were computed only by specifying map and reduce functions implemented in Hadoop. The vibrational potential energy functions were computed for at least 50 C-H stretching oscillators of the CF_3H moiety within the CF_3H - $(CH_3)_2O$ dimer at B3LYP, HF and MP2 levels of theory. The results show that there is an excellent matching between average vibrational frequencies computed from the DOS distributions, and the single vibrational frequency values computed from only a single averaged vibrational potentials. By using the results, the vibrational Schrödinger equation was solved in a variational manner. Since there are many big-data oriented problems in the scientific domains, the MapReduce paradigm and Hadoop can be suitable for their solving, especially when some reduction of the data should be performed. Our future work is oriented to solving more difficult problems in the scientific domains by using the MapReduce method and Hadoop framework.

References

[1] J. Ekanayake, S. Pallickara, and G. Fox, "Mapreduce for data intensive scientific analyses," in *Proceedings of the 2008 Fourth IEEE International Conference on eScience*, ser. ESCIENCE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 277–284. [Online]. Available: http://dx.doi.org/10.1109/eScience.2008.59

[2] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. [Online]. Available: http://doi.acm.org/10.1145/1327452.1327492

[3] C. He, "Molecular dynamics simulation based on hadoop mapreduce," Ph.D. dissertation, Computer Science and Engineering, Department of University of Nebraska-Lincoln, Lincoln, Nebraska, May 2011.

[4] C. Bunch, B. Drawert, and M. Norman, "MapScale: A Cloud Environment for Scientific Computing," University of California, Tech. Rep., Jun. 2009. [Online]. Available: http://www.google.ch/search?q=Future+of+MapReduce+for+scientific+computing&#38;ie=utf-8&#38;oe=utf-8&#38;aq=t&#38;rls=org.mozilla:de:official&#38;client=firefox-a

[5] Y. Tabaa and A. Medouri, "Towards a next generation of scientific computing in the cloud," *International Journal of Computer Science Issues*, vol. 9, no. 3, November 2012.

[6] C. Zhang, H. Sterck, A. Aboulnaga, H. Djambazian, and R. Sladek, "Case study of scientific data processing on a cloud using hadoop," in *High Performance Computing Systems and Applications*, ser. Lecture Notes in Computer Science, D. Mewhort, N. Cann, G. Slater, and T. Naughton, Eds. Springer Berlin Heidelberg, 2010, vol. 5976, pp. 400–415. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12659-8_29

[7] J. B. Buck, N. Watkins, J. LeFevre, K. Ioannidou, C. Maltzahn, N. Polyzotis, and S. Brandt, "Scihadoop: Array-based query processing in hadoop," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '11. New York, NY, USA: ACM, 2011, pp. 66:1–66:11. [Online]. Available: http://doi.acm.org/10.1145/2063384.2063473

[8] E. Kohls, A. Mishev, and L. Pejov, "Solvation of fluoroform and fluoroformdimethylether dimer in liquid krypton: A theoretical cryospectroscopic study," *The Journal of Chemical Physics*, vol. 139, no. 5, pp. –, 2013. [Online]. Available: http://scitation.aip.org/content/aip/journal/jcp/139/5/10.1063/1.4816282

[9] V. Kocevski and L. Pejov, "Anharmonic vibrational frequency shifts upon interaction of phenol(+) with the open shell ligand o2. the performance of dft methods versus mp2," *The Journal of Physical Chemistry A*, vol. 116, no. 8, pp. 1939–1949, 2012. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jp209801s

[10] R. Lmmel, "Googles mapreduce programming model revisited," *Science of Computer Programming*, vol. 70, no. 1, pp. 1 – 30, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167642307001281

[11] T. White, *Hadoop: The Definitive Guide*, 1st ed. O'Reilly Media, Inc., 2009.

[12] T. Spangler, "Algorithms for grid graphs in the mapreduce model," Master's thesis, University of Nebraska-Lincoln, 2013.

[13] H. Karloff, S. Suri, and S. Vassilvitskii, "A model of computation for mapreduce," in *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '10. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2010, pp. 938–948. [Online]. Available: http://dl.acm.org/citation.cfm?id=1873601.1873677

[14] D. Licari, "Mapreduce," November 2010.

[15] G. Wang, "Evaluating mapreduce system performance: A simulation approach," Ph.D. dissertation, Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA, August 2012.

[16] T. A. S. Foundation. Apache hadoop. [Online]. Available: http://hadoop.apache.org/

[17] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: A warehousing solution over a map-reduce framework," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1626–1629, Aug. 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1687553.1687609

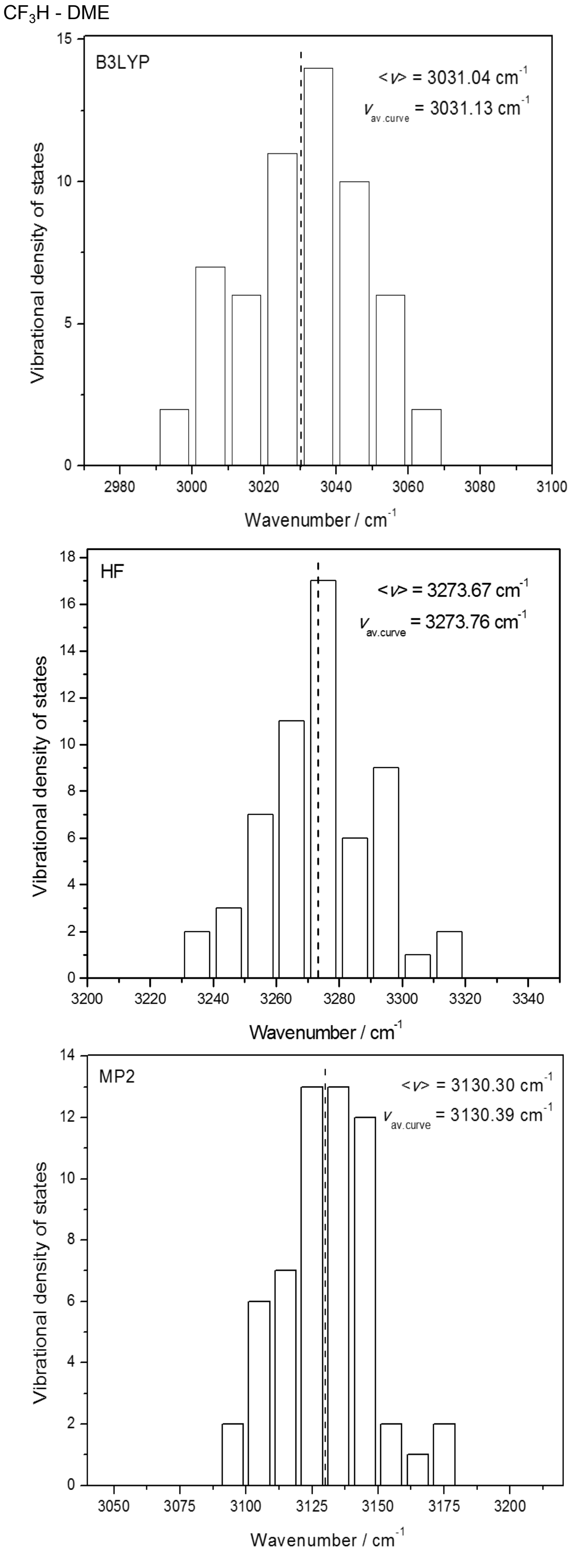


Fig. 1. Vibrational Density-of-states (DOS) Histograms Generated from the Computed Vibrational Frequencies together with the Delta-like Function

Engaging Learning Proces through Cloud Computing Models

Agon Memeti, Betim Çiço
Faculty of Contemporary Sciences and Technologies
South East European University
Tetovo, Macedonia
{am13394, b.cico}@seeu.edu.mk

***Abstract*—E-learning through cloud computing models as a new trend in education nowadays is an attractive environment for students, faculty members and researchers. Providing universities and research centers with powerful and cost-effective computational infrastructure, student's connection to campus educational services through their personal mobile devices from anywhere and faculty members efficient and flexible access to their course material in their classrooms it is what this paper is trying to resolve and discuss. Also, proposing a combination of appropriate supporting content, learner collaboration and interaction, on-line support in educating, and engaging the learning process anytime, anyplace and on a just-in-time basis.**

***Keywords*—e-learning; models; cloud computing; cost-effective; research; infrastructure.**

I. Introduction

Cloud Computing is the technology that obviously is utilizing all recent achievements in networking, distributed computing and virtualization. It allows for a greater agility and cost management of digital information's of every organization, company and especially in education through a simple flexible implementation by providing efficient access to all computer services.

Its implementation is currently much more widespread offering the ability to send and receive information's over the network using a set of functionalities.

The technology uses the Internet and central remote servers to maintain data and applications, allowing for much more efficient computing by centralizing storage, memory, processing, and bandwidth [1].

II. Cloud Service Models

Services offered by this technology are grouped into three categories:

A. IaaS (Infrastructure as a Service)

Allowing the usage of hardware computing resources as a service provider and customers to purchase hardware resources (storage, switches, routers, etc) as outsourced services, by which we can reduce the physical computing resources in a very short period.

B. PaaS (Platform as a Service)

PaaS is the relationship between SaaS and IaaS, defining the application development environment, by producing applications more quickly and with a greater degree of flexibility.

C. SaaS (Software as a Service)

Resource scaling based on demand, processing large amount of data by allowing organizations to save IT investment on infrastructure.

III. Types of Clouds

There are several models for the systems that use the paradigm of Cloud Computing. The idea is choosing the appropriate model to solve a specific problem.

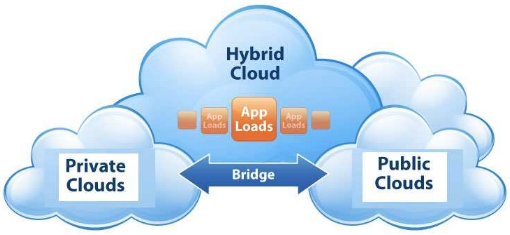


Fig.1. Cloud Computing Platforms [2]

The private cloud is established for a specific group or organization and limits access to just that group. Private clouds are built exclusively for a single enterprise. They aim to address concerns on data security and offer greater control, which is typically lacking in a public cloud [3].

The name public cloud refers to the standard model of Cloud Computing which the service is available to anyone on the Internet infrastructure (its software or hardware) free or by paying certain amount related to the volume or time of use thereof, while the Hybrid Cloud model is the combination of the two models described above so that the advantage of physical location of the information's managed by exploits private clouds with the ease of expanding public cloud resources.

IV. Cloud Computing For E-Learning

E-learning systems, also research labs usually require many hardware and software resources. These systems need to improve its infrastructure, which can devote the required computation and storage resources for these systems.

The model that we propose it will be operated for the exclusive use, enabling their own users (learners, instructors, and administrators) to perform their tasks effectively with less cost by utilizing the available cloud based application offered by the cloud service providers. The research labs in every University using this model would take advantage in terms of efficiency, reliability, portability, flexibility, and security impacts, which have been a challenging concern for cloud computing.

Setting up a private cloud has a steep learning curve for the whole institution, for both users and administrators. Although highly scalable, with possibility for installing at a number of sites, it seems that the work needed to install and maintain large stage deployments is too big. There is a need of more than average level of skills required for installing, managing, and using a private cloud [4].

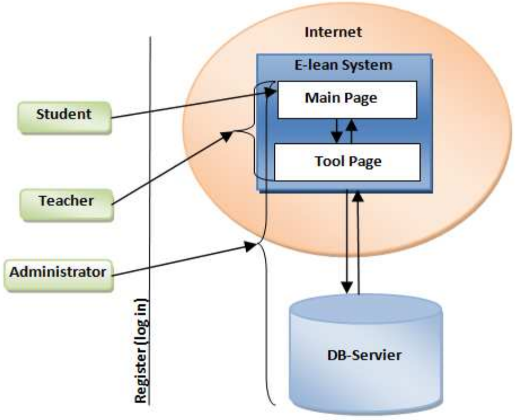


Fig.2. E-Learning Structure [5]

Our proposed model is based on Business Process Improvement (BPI) technique. The current system is operating in our intranet only, so we are improving it by private cloud platform, as teaching is often considered as the main activity in most universities. Although there are, research oriented universities or specialized research institutions, their numbers are much smaller than teaching oriented universities. In terms of teaching, computing facilities are necessary for student laboratories and libraries.

Traditional University systems are working locally based on servers and connected externally with the internet in one location inside it, but the key problem is that Project Participants are required experienced persons taking in consideration the system administration; Depreciation of Current Technology; Organizational setting (long-linked technologies are the most and hardest to change; Resources and commitment (Computer systems, space).

This would be a completely new story providing better: Mobile, decentralized and just in time learning; cost effective; speed of implementation and updating; virtualization; easy to monitor data access; latest dependency on IT department, in the meantime for some of actors it will be assessed the security still there is a private cloud.

V. Proposed Cloud Platform

In general the network topology for University labs, especially SEEU 816 Data Center is design in that manner to offer isolated services. In general the entire buildings are isolated in two logical networks: laboratories subnet and public subnet.

Main subnet network that supports laboratories environment is the core subnet and, is serving to provide base services for laboratories. This isolated subnet is very scalable and it offers redundancy network access for each laboratory in accordance to the service and specifics that each laboratory requires.

The second logical network it is the public network access subnet. In this topology design resides the public IP subnet of IT department. Using this subnet the IT can publish all services to the internet and beyond.

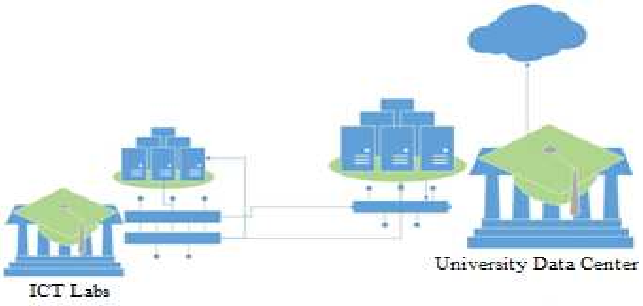


Fig.3. ICT Lab Topology

As can be seen from the logical design the ICT Lab topology, fig. 3, has to separate links from which is connected to the main University Data Centre. The first link as described above is the main subnet which connects the laboratories by using distribution layer switch with the core layer switch on the end side University main Data Centre.

This versatile virtualized platform in some way offer unified on premise environment where provisioning and automation are the main benefits (for the administration and management part) whereas self-service is what faculty staff and students benefit too.

Looking forward to make a truly private Cloud Platform it is what Universities with existing infrastructure should think about; the proposed topology is shown below in fig. 4, providing effective usage of resources, scalability, on-demand service and also a low cost solution to the 816 SEEU Data Center, starting by: Services and support to a wide range of users; A wide-range of course materials and academic support tools to instructors, teachers, professors, and other educators and university staff; Research level computational systems and services in support of the research mission of the university; Excellent resource utilization depending on different user demands; Variety of diverse service environments; Decentralized learning; The new economic normal, would bring hardware costs down [6]. Computer hardware is expensive to buy, to maintain, and to keep current.

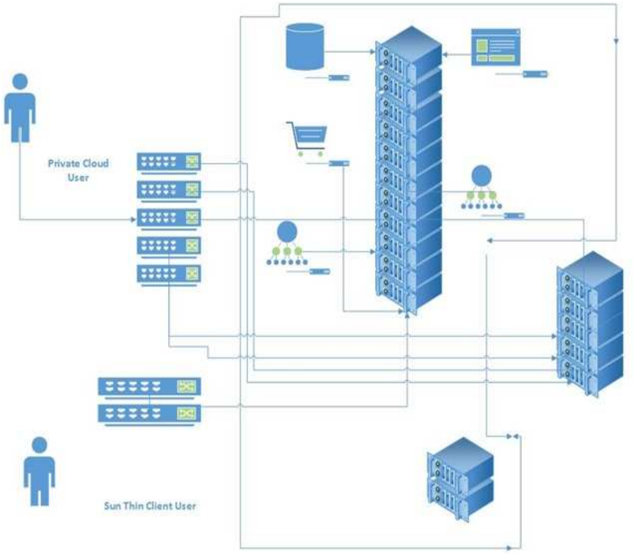


Fig.4. Proposed Labs in the Cloud

The idea is having a virtual machine inside servers in the cloud which it would have an e-Learning Platform inside it, as proposed in the model of the fig. 5:

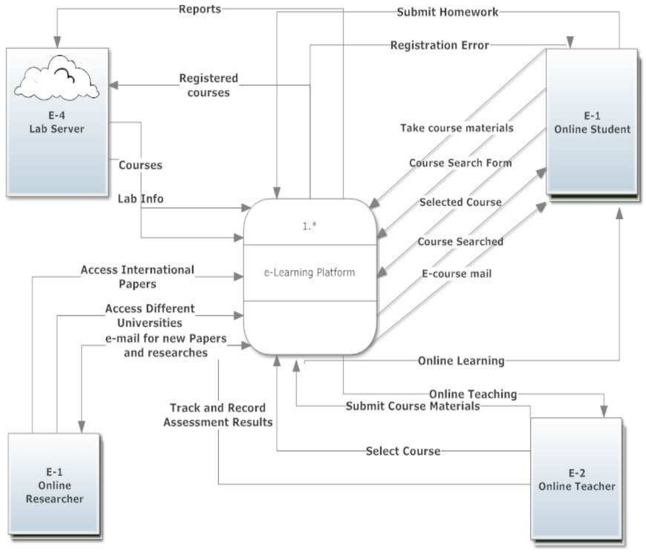


Fig.5. Proposed e-Learning Model

The proposed model should Increasing Quality and Value of E-Learning through the implementation of Cloud Computing Models, offering:

- Comfortable lab environment. Where labs should be located in a designated space, as opposed to a common area, to minimize distractions and help students focus on coursework.
- Offering sufficient internet connectivity. access online courses simultaneously, considering infrastructural bandwidth upgrades to increase network access speed.
- Supply computer accessorises and communications technologies.
- Ensure hardware and software is up to date. Administrators should work with vendors as early as possible to identify these requirements.
- Consider a range of communications options.
- Offer adequate technical support. Because computers with Internet access are critical for students taking courses in online learning labs, technical support must be adequate to keep the computers functioning properly and to solve problems that students and facilitators cannot.
- Services offered to other Universities based of agreements between Universities. Other Universities can connect different LAB services; Exchange experience on different courses held all around the world; Students can check and listen different lectures held in different universities; Also students can check all researches which are done from other students to other universities.
- Offering utilities for scientific research. Connected with European Network of Scientific Research; Grid on Cloud to increase the performance of the services.

VI. Platform Benefits

A. *Better student experience*

Is important that the present desktop virtualization for the labs that is in use to be upgraded so the students to gain a much better and performing environment that allows a much wider utilization of applications.

B. *Deployment of a Fortieth LAB*

The new upgraded server infrastructure can accommodate a networking and security lab that will enable the professors together to create a very comprehensive curriculum about security, security practices, and defense against cyber threats.

C. *Deployment of more advanced simulation solution*

The new virtual environment can accommodate new simulation solutions like*:*

- *MatLAB* - for data processing, data analysis, advanced mathematics computation, signal processing, and control systems
- *National Instruments Multisim* - for high-level electronics simulations, FPGA programming.

D. *Security research*

Safe code practices are vital for secure IT infrastructures, so by researching into advanced hacking techniques like fizzing, memory injection, and buffer overflows can enable the professors to have more data, and publish materials.

E. *High computing*

By redesigning and increasing the capabilities of the high performance, cluster is possible to perform new experiments in the grid research, high distributed computing, mathematical simulation, signal processing, computer programming algorithms research.

Conclusion

The Cloud Computing can also be used for universities, by implementing a private cloud platform with their existing infrastructure and get benefited by: accessing the file storages, educational resources, research applications, and tools anywhere for faculty, administrators, staff, students, and other users in university, on demand. The main goal of suggested prototype is: managing effectively the technological needs of universities such as delivery of software, providing of development platform, storage of data, and computing, and the most important increasing the quality and value of the University.

References

[1] A. Mansuri and P.Rathore, Cloud Computing: A New Era in the Field of Information Technology Applications and its Services, American Journal of Information Systems, 2014.

[2] Abijith Mg, Hybrid Cloud Model, P3 InfoTech. http://blog.p3infotech.in/2013/hybrid-cloud-model/ [last access on 14.05.2014]

[3] T. Harris, Cloud Computing – An Overview, SOA.

[4] P. Zoran and B. Ali Muhammad, Guidelines for Building a Private Cloud Infrastructure, ITU Technical Report, 2012.

[5] M. Adi, M. Zahraa, Cloud Computing Education, European, Mediterranean & Middle Eastern Conference on Information Systems, 2013.

[6] A. Shaban Muzhir, I. Salah Mohammed, Efficient Virtual Universities via Cloud Computing Environment, Journal of Emerging Trends in Computing and Information Sciences, 2012.

Survey of Technologies for Real Time Big Data Streams Analytic

Zirije Hasani
Faculty of Computer Science
Public University of Prizren "Ukshin Hoti"
Prizren, Kosovo
zirije.hasani@uni-prizren.com

Margita Kon-Popovska, Goran Velinov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University of Skopje
Skopje, Macedonia
{margita.kon-popovska, goran.velinov}@finki.ukim.mk

Abstract—**The beauty of web is that it empowers us with information, but challenge is that usually big amount of information has to be analyzed in real time. In the recent years the amount of data is growing in an enormous manner, some of reasons being social networks, sensor data, ATM transaction and similar. This survey deals with technologies and architectures that enable analyze of the big data in real time, commonly referred as Real Time Big Data Streams Analytic.**

In this paper we have examined several platforms: Storm and Kafka, ParStream, OpenPDC, DataTurbine, SQLStream, Hadoop Splice Machine and Apache S4 that are used for real time Big Data Stream Analytic. Our goal is to provide an overview of the most used technologies by giving a brief description for each platform their advantages and disadvantages. We also briefly review alternative technologies and architectures such as Lambda architecture that is a good solution for real-time stream analytic. Finally, we suggest Storm as the best technologies to choose among the open source platforms.

Keywords—Hadoop, Big Data streams analytic, Storm.

I. INTRODUCTION

The emerging phenomenon called Big Data is forcing numerous changes in businesses and other organization. Managing big data brings together old and new technologies and practices. The powers of one organization are information which is extracted from data. In the recent years this amount of data is growing in an enormous manner, the data are of different type and the most challenging are real time data where huge amount of data arrive in short period of time. Usually it is not possible (or there is no need) all of them to be stored and/or in the short period of time its importance will be negligible. Typically when we're talking about real-time or near real-time systems, what we mean is architectures that allow us to respond to data as it is received without necessarily persisting it to a database first [5]. That's the primary significance of the term real-time, meaning that we're processing data in the present, rather than in the future.

The process of analyzing data that came in real time is known as Stream Analytic. Stream Analytics can help firms make operational decisions, adapt to the business environments, and serve customers better when it really matters – in real-time.

In this paper we examine what Big Data Stream Analytic is, explore various technologies used and compare them in order to conclude which is the best choice for use in the case of real-time big data analytic.

The motivation to make this survey is that there is a lot of research and articles on this subject, but we did not find a survey where technologies and related architectures are compared, and recommendations for best or right choice is given for real time Big Data Stream Analytic.

In the paper we also examine and compare technologies for batch analytic. We conclude that Hadoop like systems are not appropriate platform for real-time stream analytic, Hadoop platform is a great solution for offline Big Data processing. But there are many use cases across various domains which require real-time or near real-time analytics. Real-time systems perform analytics in short time windows, i.e. correlating and predicting events based on streams generated in the last several minutes.

In section II the challenges of real time Big Data analytic process and overview of some technologies for real time stream analytic are given. In section III, technologies and related architectures are compared. As a best choice among the open source platforms Storm platform based on Lamda architecture recommended. In section IV, some of many benefits of real time Big Data stream analytic are presented. The last section is the conclusion from this survey.

II. TECHNOLOGIES USED FOR REAL TIME BIG DATA STREAMS ANALYTIC

Real Time data processing of Big Data is a complex task since it has to handle at the same time volume, variety and velocity of data. Especially to handle the velocity of data is not an easy task. Related technology should be able to collect the data generated by real time events streams coming in at a rate of millions of events per seconds. It needs to handle the parallel processing of this data as and when it is being collected. At a same time it should perform event correlation using a Complex Event Processing engine to extract the meaningful information from this moving stream. These steps should happen in a fault tolerant and distributed way [3]. The real time system should be a low latency system so that the computation can happen very fast with near real time response capabilities.

Fig. 1 illustrates a different concept of a real time system. Streaming data can be collected from various sources, processed in the stream processing engine, and then deliver the result to destination systems. In between, the Queues are used for storing/buffering the messages, [3].

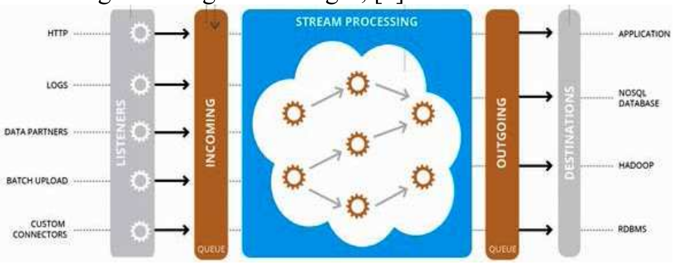


Fig 1. A different construct of a real time system. [3]

There are a large number of big data technologies many of them are open source and if the question is can we solve the Big Data problems just with open source technologies, the answer is yes because there are many open source technologies that are used by big companies. Fig. 2 lists some open source Big Data analytic platforms.

Fig. 2. Big Data analysis and platforms, [8].

Some open source big data products include [9]:

- Apache Hadoop for data storage and processing
- Apache Sqoop for bulk data transfer between Hadoop and databases
- Apache Storm, Spark, and Flume to upload and process log, event, and streaming data
- Talend Open Studio for Big Data, a graphical IDE for systematically loading all data into Hadoop, transforming, and delivering data
- NoSQL databases such as Apache Hbase, Cassandra, Couchbase, MongoDB, and Neo4j
- Analytics and BI tools such as Jaspersoft
- Cloud-based big data platforms such Cloud Foundry PaaS architecture

Next we examine following real time big data stream analytic technologies:

A. *Storm and Kafka*

Storm is a free and open source distributed real-time reliable and fault-tolerant system for processing streams of data, [11]. It is simple and can be used with any programming language. Born inside of Twitter, Storm does for real-time processing what Hadoop did for batch processing.

Kafka for its part is a messaging system developed at LinkedIn to serve as the foundation for their activity stream and the data processing pipeline behind it.

When paired together, Storm and Kafka enable stream processing at linear scale, assured that every message gets processed reliably in real-time with data velocities rate of tens of thousands of messages every second. Additional advantage is their superior approach to ETL (extract, transform, and load) and data integration [3], in-memory analytics and decision support in real-time.

As such, Storm and Kafka are already in use at a number of high-profile companies including Groupon, Alibaba, and The Weather Channel. Companies are quickly realizing that batch processing in Hadoop does not support real-time business needs.

B. *ParStream*

The ParStream is used in many industries as the real-time big data analytics platform that delivers ultra-fast interactive analytical results on massive structured and semi-structured data, [7]. It provides sub-second response times on billions of data records while continuously importing new data.

ParStream provides the ability to

- Analyze and filter billions of records
- Query data structures with 1000's of columns
- Get answers in milliseconds without Cubes
- Continuously import data with low latency
- Execute 1000's of concurrent queries

ParStream runs on standard infrastructure – on single servers, dedicated server-clusters and in private and public clouds. Most Linux-distributions are supported, including RHEL and SLEs [7].

ParStream was specifically engineered to handle [7]:

- Structured and semi-structured data
- Denormalized, very large fact tables
- Selective and multi-dimensional filtering and analytics
- Very short query response times
- Very high query throughput

C. *OpenPDC*

The OpenPDC is used to manage process and respond to dynamic changes in fast moving streaming phasor data, [10]. More specifically, the openPDC can process any kind of data that can be described as "time-stamped measured values". These measured values are simply numeric quantities that have been acquired at a source device and are typically called points, signals, events, time-series values or measurements. The type of data that can be analyzed are from many area some of them are: consumer energy usage (smart-grid), seismic metering, high-speed location tracking, fast changing temperature monitoring, surveillance applications, network traffic processing, etc.

D. DataTurbine

The DataTurbine transitioned in 2007 year from commercial to open source system under the Apache 2.0 license. It is a buffered middleware, not simply a publish/subscribe system. It can receive data from various sources (experiments, web cams, etc) and send data to various sinks (visualization interfaces, analysis tools, databases, etc). It has "TiVO" like functionality that lets applications pause and rewind live streaming data. DataTurbine lets you stream data and see it in real-time, but it also lets you TiVo through old and new data, share it with anyone over the network, do real-time processing of the streams and more.

The data type that can be analyzed with DataTurbine are: stream live data from experiments, labs, web cams and even Java enabled cell phones (weather data, load readings from a bridge, pictures from a security camera, GPS-tagged biometrics from a tracked tiger, chlorophyll readings from a lake buoy, etc.)

E. SQLstream

The SQLstream turns streaming high volume, high velocity, structured and unstructured Big Data into real-time value, [12]. SQLstream is the only platform that offers streaming integration with time-based and location-based in-memory analytics from the live machine data as they stream past. SQLstream integrates seamlessly with Hadoop and other popular high performance data warehouses. Log file, sensor, network and service data are collected in real-time, analyzed ‘on the wire’ and then streamed continuously to real-time dashboards as well as Big Data stores for further analysis.

SQLstream have a connector for Hadoop HBase. It provides continuous integration with Hadoop HBase, enabling organizations to exploit the value of both real-time analytics over the arriving data, and previously stored data and streaming intelligence in Hadoop HBase for further analysis, [12].

F. Hadoop Splice Machine

The Splice Machine is a standard SQL database, supporting real-time updates and transactions implemented on the scalable, Hadoop distributed computing platform. Designed to meet the needs of real-time, data-driven businesses, Splice Machine is the only transactional SQL-on-Hadoop database. Like Oracle and MySQL, it is a general-purpose database that can handle operational (OLTP) or analytical (OLAP) workloads, but can also scale out cost effectively on inexpensive commodity servers, [13]. Splice Machine joins two proven technology stacks: Apache Derby, a Javabased, full-featured ANSI SQL database, and HBase/Hadoop, the leading platforms for distributed computing.

G. Apache S4

The Apache S4 platform for processing continuous data streams is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.

III. WHICH TECHNOLOGIES AND ARCHITECTURE TO CHOOSE

Having available a big choice of technologies for Big Data stream analytic, question arise which one is suitable for our needs. In the Table 1 below attempt is made to extract some typical characteristics for technologies comprised in the survey. All of them are real time big data stream analytic systems and process streams of data. From the comparison table we can see that the most of the characteristics are part of Storm and Kafka combined system.

Table 1. Comparison of real time big data stream analytic technologies

Characteristics		**Platforms for real-time Big Data stream analytic**						
	Storm	***Kafka***	***ParStream***	***OpenPDC***	***DataTurbine***	***SQLStream***	***Hadoop Splice Machine***	***Apache S4***
Open Source	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Robust, Foult tolerant	Yes	Yes	Yes		Yes		Yes	Yes
Scalable	Yes	Yes			Yes		Yes	Yes
Ad hoc queries	Yes			Yes				
Distributed	Yes	Yes	Yes	Yes			Yes	Yes
In-memory Analytic	Yes		Yes			Yes		
Suport SQL						Yes	Yes	
Structured data			Yes			Yes		
Semi-structured data			Yes					

Characteristics		Platforms for real-time Big Data stream analytic						
	Storm	*Kafka*	*ParStream*	*OpenPDC*	*DataTurbine*	*SQLStream*	*Hadoop Splice Machine*	*Apache S4*
Unstructured Data						Yes		
Data from instruments				Yes	Yes			
Data from social networks	Yes	Yes				Yes		
Company owned	Twitter	Apache						Apache Incubator project
Capacity for asnalytic	A million tuples processed per second per node	Hundreds of megabytes of reads and writes per second from thousands of clients	Analyze and filter billions of records					Process thousands of search queries per second
Hybrid On-Disk and In-Memory Storage			Yes					
Stored data	No		Yes			No		
Paralelization			Yes					Yes

Another interesting question is choice of suitable architecture. An interesting proposal is the Lambda architecture of Nathan Marz [14]. The Lambda Architecture shown in Fig.3 solves the problem of computing arbitrary functions, on arbitrary data, in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Kafka-Storm[1] for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general and extensible, allows ad hoc queries, need minimal maintenance and is debuggable. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop is doing for batch processing.

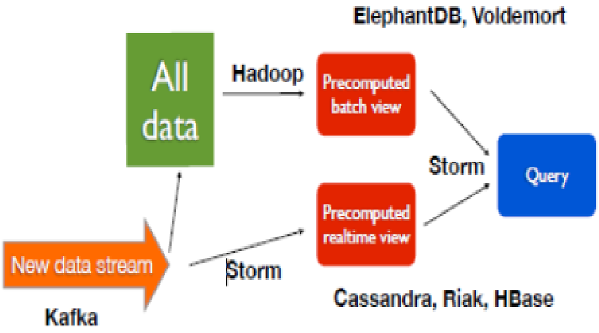


Fig. 3. Lambda architecture

SAMOA [15] which is shown in Fig. 4 is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.

IV. Benefits from real time Big Data stream analytic

Thinking about Big Data Stream Analytics the first thing we imagine is a big space where we have to store the data. Immediately question arises is there a need to store all this data and is it possible to store all of them. Starting from the known capacity of Big Data systems the answer is really simple we need to store and to analyze a small part of this data which are of our interest.

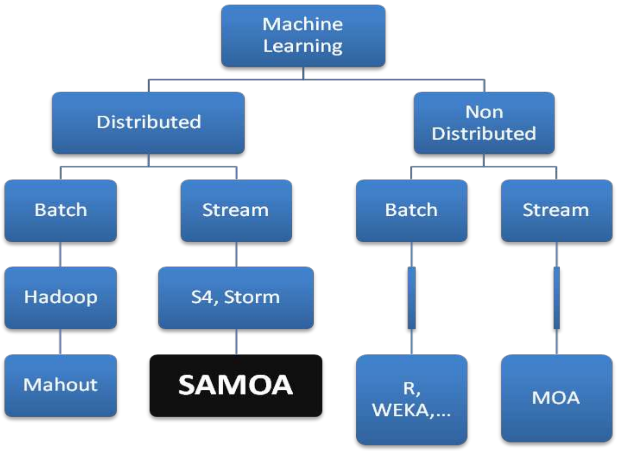


Fig. 4. SAMOA architecture

The difference between data stored before analytic and real time data is that if data is even one day old, the insights of data may already be obsolete. Companies need to analyze data in near-real-time, often in seconds.

Traditional data quality batch processing is no longer enough to fully sustain effective operational decision-making. Integrating, cleansing and analyzing data in real-time allows a company to engage in opportunities instantly. For example,

[1] http://storm-project.net/

using real-time data processing, a company can personalize a customer's on-line website visit, enhancing the overall customer experience, can make instant fraud detection, can monitor huge sensor networks and similar.

There are many benefits from real time Big Data stream analytic, the research in [2] show some of this benefits. The results from the research are shown in Fig.5 below.

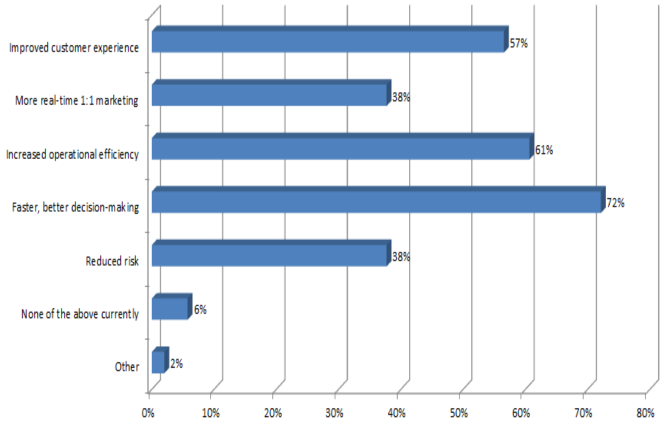


Fig. 5. Anticipated benefits from Real-time Big Data Analytics, [2].

The survey respondents were asked to weigh in on the anticipated benefits from real-time Big Data Analytics. 72% ranked faster, better decision-making as a result of real-time analytics as the top benefit. Increasing operational efficiency ranked #2 with 61% of the respondents indicating it as a major benefit. Improving customers' experience through real-time analytics was a close third at 57%. Interestingly, real-time 1:1 marketing tied at #4 along with risk management suggesting that few organizations have really explored or understood the true potential of real-time Big Data Analytics for increasing revenue, [2].

From this study there is a conclusion that technology support for streaming Big Data Analytics and action is limited. However, 67% of those surveyed indicated that their organizations had little to no technology support for analyzing streaming Big Data to take immediate action, [2]. The results are shown in Fig. 6.

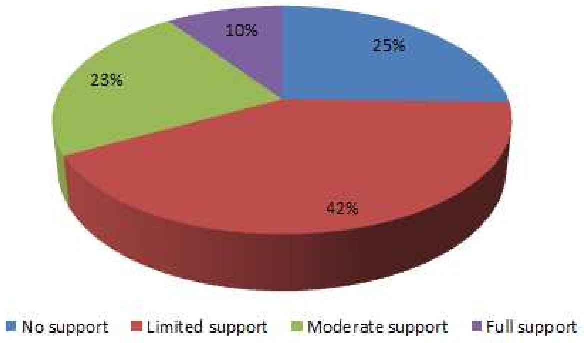


Fig. 6. Support for analyzing streaming Big Data for immediate action, [2].

V. Conclusion

We conclude that Storm and Kafka are the best choice for real time big data stream analytic. This conclusion came from the fact that the most of characteristics compared in this paper are owned by these two systems. Storm and Kafka are open source systems also they allow in memory analytic and ad-hoc query, they are robust, scalable and fault tolerant too. This new systems have inhered the characteristics from the old platform which are used for analyzing small amount of data.

Storm is proposed also by Lambda architecture which solves the problem of computing arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer.

The idea of real time big data stream analytic system is to combine different systems and also to blend the process of batch analytic and real time stream analytic. For example Lambda architecture is implementing this idea which brings together Hadoop, Storm and Kafka. Also, there are many platforms for real time Big Data analytic and all of them face with the same challenge analyzing data at the moment they are coming because letter they are not important. There is a 10% of full support for analyzing streaming Big Data for immediate action, which is a small percentage.

References

[1] A. Bifet, "Mining Big Data in Real Time," Yahoo! Research Barcelona, Spain Informatica 37, pp. 15-20, December 15, 2012.

[2] Vitrina, "The State of Real-time Big Data Analytic: 2013 Survey Results," Octomber, 2013.

[3] D. Bhattacharya, M. Mitra, "Analytics on Big Fast Data using Real Time stream data processing architecture," EMC, 2013.

[4] D. Bonino and L. Russis. "Mastering Real-Time Big Data With Stream Processing Chains," XRDS, VOL.19, NO.1, 2012.

[5] M. Barlo, "Real-Time Big Data Analytic: Emerging Architecture," O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA. 2013.

[6] Doug Henschen. *16 Top Big Data Analytics Platforms*, (2014, January). Retrieved February 15, 2014, from http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609.

[7] ParStream. *The Real-time Database for Big Data Analytics,* (2013). Retrieved March 10, 2014, from https://www.parstream.com/product/.

[8] Big Data Startup. *The Big Data Open Source Tools*, (2014). Retrieved March 10, 2014, from http://www.bigdata-startups.com/open-source-tools/

[9] F. Halper, "Big Data Management: Governance in a changing data world," TDWI E-Book, January 2014.

[10] OpenPDC: Grid Protecting Allience. *OpenPDC*, (2014). Retrieved February 15, 2014, from http://openpdc.codeplex.com/releases

[11] Apache. *Storm, distributed and fault-tolerant realtime computation*, (2014). Retrieved February 24, 2014, from http://storm-project.net/

[12] SqlStream. *Big Data Analytics,* (2014). Retrieved March 03, 2014, from http://www.sqlstream.com/solutions/big-data-analytics/.

[13] Hadoop Splice Machine. *The Real-Time SQL-on-Hadoop Database,* (2014). Retrieved March 08, 2014, from www.splicemachine.com

[14] N. Marz and J.Warren. Big Data: Principles and best practices of scalable realtime data systems. ManningPublications, 2013.

[15] SAMOA. *Scalable Advanced Massive Online Analysis*, (2014). Retrieved March 20, 2014, from http://yahoo.github.io/samoa/.

Session 11

Best Student Paper Competition

The IEEE Computer Chapter Macedonian Section organized a best student paper competition. In the student session 13 student projects were presented, and the best 3 were awarded. The awarded student authors have been encouraged to submit their papers to relevant journals and offered help in the process of doing so.

- 1st prize: Aleksandar Ristovski
  Nested Parallelism Concepts of Ray Tracing Algorithms and Multithreading API Performance Analysis
- 2nd prize: Aleksandar A. Trposki
  Use of grid computing to solve matrix based games of medium size solving bubble blaster
- 3rd prize: Aleksandar Stojanovski
  Unconventional interaction with computer games (BikeSimulator)

Figure 11.1: IEEE Best Student Paper Award

ISBN 978-608-4699-04-0

9 786084 699040